Investigating Protein Evolution Through Sequence Space Using a Biophysical Lens

by

Michael James Shavlik

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Biology

Dissertation Committee:

Dr. Matthew Barber, Chair

Dr. Michael Harms, Advisor

Dr. Nadia Singh, Core Member

Dr. Scott Hansen, Core Member

Dr. Raghuveer Parthasarathy, Institutional Representative

University of Oregon

Fall 2023

DISSERTATION ABSTRACT

Michael James Shavlik

Doctor of Philosophy in Biology

Title: Investigating Protein Evolution Through Sequence Space Using a Biophysical Lens

How do the underlying biophysical properties of proteins dictate the "rules" that govern molecular evolution? Understanding the principles and mechanisms that determine which evolutionary trajectories proteins take is crucial to protecting humans against viral protein evolution and developing therapeutic, custom, drugs through protein engineering. Although many approaches have been developed to investigate the process of protein evolution, a deep understanding of the relationship between sequence space and protein biophysics can alleviate key deficiencies in our knowledge. What is the underlying distribution of functional proteins in sequence? Do specific biophysical properties dictate the interconnectedness of these functional proteins? How does the protein energy landscape change across evolutionary time and how can that inform our understanding of evolution? This dissertation will explore two methods of answering these questions: 1) High-throughput mutagenesis and phenotype characterization to explore sequence space using fluorescent proteins and 2) Ancestral Sequence Reconstruction linked to a biophysical lens using protein energy landscapes.

This dissertation includes previously published material.

CURRICULUM VITAE

NAME OF AUTHOR: Michael James Shavlik

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR

Doane University, Crete, NE

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2023, University of Oregon

Bachelor of Science, Biology, 2018, Doane University

AREAS OF SPECIAL INTEREST:

Evolutionary Biology

Biophysics

Molecular Evolution

PUBLICATIONS:

Shavlik MJ, Harms MJ, Fitzgerald B, Kotamarti A, Chisholm, L, Marquesee S, Petersen M., Local Neighborhoods in GFP-Like Protein Sequence Space Change over Evolutionary Time (*in preparation*)

Garcia, AC, **Shavlik MJ**, Harms MJ, Pluth MD., Automatically Measuring Ellipticity of Small Molecule CBn Structures (*in preparation*)

Chisholm LO, Orlandi KN, Phillips SR, **Shavlik MJ**, Harms MJ., Ancestral reconstruction for biophysics *Annual Reviews in Biophysics* (2024)

ACKNOWLEDGEMENTS

I would like to extend a special thanks to my advisor, Mike Harms, for all his keen guidance, support, and mentorship throughout my time in graduate school. Without his help, none of this work would have been possible. I also want to acknowledge the early-career support given to me by my past advisors before graduate school: Erin Doyle, Tessa Durham Brooks, and Jay Hollick. I would also like to thank the other faculty that supported my endeavors in graduate school – Laurel Pfiefer-Meister for supporting my teaching career while conducting this research, and Jeff Bishop for overseeing and supporting all my cytometry experiments. The rest of the Harms lab members have been a wonderful support system and I'm grateful to each of them for always being available to bounce new ideas off each other and mutually grow as scientists. Finally, a massive thanks to my family, friends, and wife, Ry, for all their unending love and support.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Figure Page

CHAPTER I

INTRODUCTION

Mutations alter the functions of proteins

Proteins are biological macromolecules that are responsible for performing an innumerable array of specific and diverse functions across all taxa of life. Prokaryotes, Eukaryotes, and Archaea alike all use the powerful functional diversity of proteins to sustain important physiological processes[1–8]. Fortifying cellular structure[9,10], processing molecules of energy[11], and breaking down harmful toxins[12] are just a few critical roles that proteins fill in upholding organismal homeostasis.

While proteins are crucial for maintaining these key life processes, they can often be surprisingly mutable in both structure and function[13–17]. Stretches of DNA that encode proteins are subject to spontaneous mutations through many means, such as replication errors, gene duplications, and transposable element translocations[18–20]. These mutations, which are introduced on the DNA level, can have several potential effects on the downstream translated protein. These effects can range from having nearly no noticeable effect on a protein's structure or function[21], to complete disruption of 3-dimensional structure and ablation of function[22], and everything in between. While the relative probability of a neutral mutation to a deleterious one depends heavily on several factors, such as mutation location, mutation type (e.g. substitution vs frame-shift), and

the function of the protein, most mutations are ultimately expected to be have a disruptive and negative impact on protein structure and function[23].

Mutations can also either improve an existing function or completely change the function of a protein[24,25]. In some cases, a single amino acid mutation can be responsible for changing protein function[26], while in other scenarios, multiple, small-effect mutations may sequentially lead to a new function of the protein[27]. This process – proteins acquiring mutations over time that may lead to functional changes – is the process of protein evolution.

It is important to study the process of protein evolution

There are many angles that have been taken to study this protein evolution, ranging from theoretical studies investigating the nature of the evolutionary process itself, all the way to explicitly practical methods of exploiting protein evolution in the lab to engineer better proteins or improve human health.

Protein engineering is a field that seeks to produce novel proteins or modify existing proteins to improve methods in the biotechnology field. This process of engineering better proteins takes advantage of knowledge of how certain mutations will affect protein structure and function, which often relies on predictive models based on previous evolutionary information[28]. Thus, a better understanding of mutational effects through evolutionary studies can improve our design of better proteins for biotechnology applications such as the generation of more efficient plastic degrading enzymes[29].

A thorough understanding of the evolutionary process itself will also lead to great human health benefits. For example, the evolution of viruses each flu season creates a public health

predicament where vaccine development depends heavily on our ability to predict what influenza strains will be most likely to appear[30]. This prediction relies on evolutionary models to statistically infer likely candidate strains on a given season based on previous data. According to the CDC, yearly influenza vaccines have been shown to be no greater than 60% effective[31,32]. Proper defense against viral evolution thus requires more refined predictive models of evolution, which foundationally requires a deeper understanding of the evolutionary process itself. Thus, a better theoretical understanding of the 'how' and 'why' proteins evolve is essential to important applications for human health and developing crucial tools in biotechnology.

Sequence space

"Sequence space" is a particularly useful framework with which to gain a better grasp of how and why proteins evolve. Protein sequence space can be defined simply as the entire set of amino acid sequences that are hypothetically possible for a given protein. In this space, a single "step" in space from one sequence to another corresponds to one amino acid change. As proteins evolve through acquiring mutations, they navigate their sequence space, moving from one possible sequence to another. Each amino acid has a set of 19 other possibilities at a given position in the primary sequence of the protein. Unsurprisingly, the entirety of protein sequence space is intractably large (e.g. a 5 amino acid long protein has a sequence space of 20^5 = 3,200,000 possible sequence variants). Considering that the average Eukaryotic protein is ~472 amino acids in length, exhaustively characterizing this entire space is an impossible task for the foreseeable future. Although the number of hypothetical sequences is massive, the number of functional sequences is extremely small in comparison[33], and the further number of sequences that are likely to be sampled by evolution is even smaller, given that mutations tend to be very

rare events. Thus, although sequence space is hypothetically huge, the amount of space relevant to natural protein evolution is quite tractable for experimental and computational studies investigating principles guiding the evolutionary process on the molecular scale.

One useful insight that these studies of protein sequence space gain is the contribution of underlying biophysical features that dictate which paths of evolution are possible and why. As proteins acquire mutations and move through sequence space, they must maintain a delicate balance of thermodynamic stability, proper folding rates, and functionality, which mutations tend to disrupt in some way[34–36]. Across all of these, biochemical properties such as electrostatic interactions, hydrophobic residue composition, and charge networks, which can all impact stability, kinetics, and function, provide useful information as to why a protein would evolve in one way and not another[37,38]. Investigating biophysical properties that underlie protein sequence space is key to understanding the functional layout of protein sequence space (what regions are available for evolution to access or not) and what features dictate that layout. Thus, investigating protein evolution through the lenses of sequence space and biophysics together can provide unprecedented insight into the evolutionary process and reveal why proteins evolve the way that they do.

Outstanding questions in the field of protein evolution and sequence space

How do we study protein sequence space? One way we can accomplish this is by uncovering the interconnectivity and organization of non-functional and functional regions of sequence space. We can use the historical evolutionary trajectory of a protein that naturally evolved a new function to study the organization of functional sequence space. Analyzing the

space around the historical trajectory is advantageous because this is a small, relevant region of sequence space that could reveal potential alternative evolutionary paths. If we can determine the likely original path that evolution took for a given protein, we can then ask how many alternative paths may have existed in nearby sequence space as that protein evolved. Was the historical trajectory through sequence space the only available path to the new function? Or were there many "might-have-been" paths that could have occurred? Intertwined with this idea, what kinds of biophysical principles allowed the historical trajectory to happen and arrive at this new function, and do those same principles open up or close off alternative paths through sequence space?

Across sequence space, we can also consider how the energy landscape of a protein changes. The energy landscape is a way to envision how stable or unstable the ensemble of structures is for a given protein. Proteins occupy a set of possible structures as they go from unfolded states to folded states, and some of these folded states are more stable energetically than others[39]. As a protein evolves, one folded state of the protein may change to become more or less likely to be populated than another because mutations can stabilize or destabilize certain structures[40]. There are many interesting questions about the relationship between the underlying protein energy landscape and the historical trajectory a protein took along its evolutionary transition. How does the energy landscape shape historical trajectories through protein evolution? Does that energy landscape change dramatically over time, or are the changes more subtle? Does the energy landscape also dictate what kinds of paths through evolution are accessible or inaccessible? Considering proteins as ensembles of structures and their underlying free energies can help us achieve a more comprehensive view of protein evolution beyond static, singular structures, which could in turn help us better understand the effects of subtle mutations.

Additionally, we can explore how biophysics might drive certain evolutionary characteristics for proteins. For example, some proteins may have high mutational robustness, or high mutational tolerance, while others may have low robustness. This mutational robustness increases the likelihood that a mutational step in sequence space will yield a functional protein and potentially allow for development of a new function. What biophysical features of a protein drive mutational robustness, and can that robustness change over evolutionary time as the biophysics change for the protein (e.g. stability, kinetics, etc.)? Evolvability is another characteristic of evolution that may be driven by underlying biophysics. Evolvability of a protein refers to the accessibility and pervasiveness of new functions in a protein's immediate sequence space. If a protein has relatively high evolvability, mutations have a higher likelihood of generating a protein variant with a new function. Are evolvability and robustness intrinsically linked properties? Are they conflicting properties, where a protein cannot be both robust and evolvable at the same time, or are they complementing properties? Do the biophysics that underlie high mutational robustness also underlie high evolvability? Finding the answers to these questions can help us not only understand the "why" of protein evolution, but also aid in efforts to engineer mutationally robust proteins for medical applications and other contexts.

One interesting and less explored avenue with robustness and evolvability and their underlying biophysics in evolution relates to how these properties might change across a single historical trajectory. As a protein acquires mutations over time, does its mutational robustness and evolvability significantly change, or are they somewhat constant features across evolution? If they do change over evolutionary time, is a single mutation sufficient for altering these features, or does it take many mutations to disrupt or improve them? Understanding the answers to these questions can provide valuable insight into tracking and preventing rapidly evolving

proteins such as viral proteins that could yield quick changes to the pathogen's virulence and severity. Models for predicting viral protein evolution on such a rapid scale might benefit from information on mutational effects on immediate robustness and evolvability changes.

Existing approaches to studying protein evolution through sequence space

While these are all outstanding questions in the field, several promising methods to address them have been developed both experimentally and computationally. One of the most widely used is Deep Mutational Scanning (DMS). In deep mutational scanning, every single point mutant is generated experimentally or computationally around a wild type protein sequence of interest. While this has been previously limited to small peptide regions of only a few amino acids due to experimental limitations[41,42], newer approaches to DMS have recently been innovated to expand the tractable sequence length to entire proteins[43,44]. Through this method, one gets detailed information about all protein variants within one mutational step in sequence space for a protein. This method, while useful for immediate information in sequence space, has several drawbacks. The main one being that any information of double mutants or beyond is totally lost and doesn't provide any insight further into sequence space. Studies have shown that epistasis, or non-additive effects of multiple mutations in a protein, is pervasive and makes evolution unpredictable. Because of this, results from a deep mutational scanning experiment cannot be extrapolated reliably beyond single mutations, which is often a goal in evolutionary biology and the characterization of sequence space.

On the computational side, as the field of machine learning has become popular for diverse applications in the world, so to has it been used in protein sequence space exploration.

While there are too many algorithms to cover succinctly, one promising class of machine learning methods is generative modeling. Generative models are intricate models that take a data set of observed and characterized inputs and learns patterns to generate an unobserved variant with a predicted value or trait[45,46]. These models have been applied to protein sequence space – learning small datasets of previously characterized protein variants with a given phenotype value and then attempting to predict phenotypes of unseen sequences elsewhere in sequence space. While some models, such as Variational AutoEncoders (VAE) have seen some success in the field[47,48], epistasis is still a large limitation. Experimental validation of some of these sequences that are subject to higher-order (3+ mutations) epistasis have shown that the predicted phenotypes and observed phenotypes often do not match[47]. Thus, predicting phenotypes for sequences far away in protein sequence space solely through computational means is not yet refined enough to be a reliable measure for answering all of these aforementioned questions, despite some promising progress.

Another experimental technique that can explore volumes of sequence space beyond a single mutational step is the coupling of random mutagenesis with high-throughput phenotype characterization. In this approach, one sparsely samples the volume of sequence space around a genotype of interest using a random mutagenesis technique such as error-prone PCR. While complete coverage of genotypes is not possible in this volume, it is possible to sample enough genotypes to acquire a general idea of the distribution of phenotypes in the space. An advantage of this method is that mutant generation is simple, quick, and can yield large and unbiased libraries of mutant proteins. This technique is especially effective when linked to high-throughput characterization techniques of phenotypes, such as fluorescent protein signal via flow cytometry[49]. Additionally, this type of approach could be applied to multiple genotypes along a

historical evolutionary trajectory to identify changes in things like mutational robustness and evolvability over time. Identifying these changes then opens avenues for biophysical testing of interesting protein intermediates along the trajectory, thus linking evolution and biophysics in a compelling way.

Another useful approach is Ancestral Sequence Reconstruction (ASR). ASR is a statistical technique that uses a dataset of evolutionarily related sequences of an extant protein of interest to backtrack in evolutionary time and infer an ancestral sequence[50]. This technique is especially helpful in recreating the likely historical trajectory of evolution for a given protein because statistical models can then be used to infer evolutionary intermediates between the ancient and extant proteins. From there, experimental characterization can be conducted in lab to identify the biophysical characteristics that changed across time for the protein. The importance of energy landscapes and their relationship to driving protein evolution can be investigated through ASR. Taking a biophysical perspective through energy landscapes when resurrecting ancient proteins and inferring their historical evolutionary path can help reveal useful links between biophysical and biochemical properties that may have dictated a protein's trajectory. The energy landscape is particularly interesting because it allows us to consider the entire ensemble of states that exist for a protein and how that dynamic interchange of possible structures plays into evolution.

Altogether, protein evolution is an extremely important process for adaptation that creates novelty in molecular form and function, which can be used to progress human health interests. A solid understanding of these principles guiding the process of evolution is crucial to effective medical applications, which can potentially be accomplished through a deeper exploration of protein sequence space and the biophysical properties underlying that space. While several

promising techniques have been developed to accomplish this exploration, this dissertation will focus on highlighting two approaches that show particular promise in their application: coupling random mutagenesis with high-throughput characterization of volumes of sequence space, and using Ancestral Sequence Reconstruction through the lens of energy landscapes.

Sampling small volumes of sequence space for a naturally evolved GFP-like protein

While other studies have used random mutagenesis couple with experimental characterization of mutant phenotypes in sequence space[49], there hasn't been much work done to implement this technique across an evolutionary trajectory of a naturally evolved protein. It is important to characterize small volumes of sequence space, or "local neighborhoods", across a real evolutionary trajectory because it can provide a dynamic view of how different protein features such as mutational robustness can change through a natural evolutionary path. To investigate local protein neighborhoods across evolution, we used a naturally evolved GFP-like protein that evolved from an ancestral green color to a photoconvertible green-red color. This protein was isolated from the great star coral *Montastraea cavernosa*, and its likely evolutionary trajectory was previously characterized by the Matz group[51]. This trajectory was inferred to likely have taken a minimal 12 mutations to reach the derived green-red photoconvertible phenotype, providing a wealth of intermediates around which to generate local neighborhoods in our study[51,52]. Additionally, mutations in this space show interesting epistatic effects. For example, a Q to H mutation at position 62 in the protein was required to develop the photoconvertible phenotype, but is insufficient on its own for conferring the phenotype when introduced into the ancestral background[51,53].

Linking both random mutagenesis and flow cytometry for fluorescence characterization of those mutants, we generated the local neighborhoods of both the ancestral and derived states of our GFP-like protein, as well as several point mutants that will be expanded upon in chapter 2. We found that local neighborhoods change dramatically across the evolutionary trajectory of this protein in both mutational robustness and evolvability. The ancestral GFP like protein was very robust to mutations, but didn't show any color diversity in its local neighborhood, while the evolved protein was much more susceptible to losing color through mutations but showed more color diversity in reds, oranges, and yellows. Additionally, we also found that a single point mutation can significantly change the mutational robustness and accessibility of a local neighborhood along the trajectory. These two results highlight the importance of characterizing volumes of sequence space across a trajectory as there can be huge shifts in sequence accessibility after a single mutation in space, which could be important considerations in predictive models of evolution.

Unexpectedly, we found that global protein stability was not the sole biophysical determinant of mutational robustness in the generated local neighborhoods. This result underscores the complex nature of protein evolution and suggests that there are likely other biophysical features of the proteins that may change across the trajectory and regulate mutational robustness. We have started to find preliminary evidence that local stability may be a better determinant of robustness changes across the evolutionary trajectory and could be an important feature when considering future evolutionary models and points of interest for protein engineering studies.

Considering the power of ASR through the lens of protein energy landscapes

Proteins are complex molecules. Contrary to the introductory biology teaching of proteins existing as a singular and static folded structure, proteins exist as an ensemble of structures, often fluctuating back and forth between conformations in this ensemble[39,54–56]. The likelihood that a protein is in a given conformation at a particular timepoint is dependent on the associated free energy of that conformation. Thus, the unfolded conformation of a protein, as well as every folded conformation has some free energy associated with it, indicating the favorability (stability) of that conformation. This concept, a protein's ensemble of conformations and their associated free energies, is often referred to as a protein's energy landscape. Understanding a protein's energy landscape can be crucial to understanding the development and complexity of protein structure and function.

Historically, evolutionary studies focus on the evolution and development of the most stable protein conformation. This is the most straightforward way to study protein evolution, as it eliminates the consideration of dynamic conformational switches in evolution and can distill mutational effects on a singular structure and its function. Investigating and considering the entire ensemble of conformations for a protein and how the free energy landscape changes is important, however, as significant shifts in this energy landscape could lead to a complete shift in the favored stable conformation, which could lead to functional changes as well. These changes in function are indirectly linked to the mutations acquired through evolution, as it is these mutations that shifted which conformation was now the most stable and favored state, thus impacting the associated function. Tracing the evolution of the ensemble and the free energy landscape is a difficult task, as there are many "moving parts" to characterize across each intermediate evolutionary step.

One method that we propose can help alleviate some of this difficulty in studying the evolution of free energy landscapes and ensembles is Ancestral Sequence Reconstruction. ASR, as mentioned earlier, is a technique that can infer and resurrect an ancient protein, and can even be used to then estimate the evolutionary intermediates between the ancient and evolved states. When coupled with biophysical characterization of these intermediate proteins, we believe that ASR can be a powerful tool in disentangling the evolution of conformational ensembles and the energy landscape. In chapter 3, we will review existing literature that highlights the versatility of ASR and propose future possibilities to integrate the technique with biophysical characterization and reveal the evolution of the free energy landscape.

In summary, this dissertation will focus on two approaches to investigating protein evolution through sequence space: generating local neighborhoods across an evolutionary trajectory and using ASR to gain insight on changes to the energy landscape over time. These two approaches will pay specific attention to the interplay of biophysics and evolution and explore the importance of experimental characterization through biophysical approaches in evolution studies. We will start with GFP-like protein neighborhood characterization in chapter 2, followed by reviewing ASR through a lens of energy landscapes in chapter 3.

Bridge to chapter II

In chapter I, I introduced the key concepts of protein evolution, sequence space, biophysics, and their connections to one another. To understand the process of protein evolution, we need to gain a deeper understanding of how proteins navigate their sequence space and characterize the change in biophysical traits across that sequence space. There are many approaches that have been recently explored for gaining that deeper understanding of protein evolution, including computational methods using machine learning and experimental methods using deep mutational scanning. In chapter II, I present an in depth study using another experimental method outlined in chapter I. Coupling high-throughput mutagenesis with fluorescence and biophysical characterization of a naturally evolved fluorescent protein, chapter II will introduce this experimental framework, the results of applying this framework, and provide considerations for future work in this area.

CHAPTER II

SINGLE SUBSTITUTIONS ALTER THE ACCESSIBILITY OF COLOR IN THE LOCAL SEQUENCE SPACES OF EVOLVING GFP-LIKE PROTEINS

Author Contributions

Michael Shavlik and Michael Harms conceptualized the study and designed experiments. Michael Harms acquired funding for the study. Michael Shavlik, Mark Petersen, Brennan Fitzgerald, Amelia Kotamarti, and Lauren Chisholm conducted the experiments and data analysis. Michael Harms administered the project. Susan Marquesee oversaw the experiments by Mark Petersen. Michael Shavlik constructed the figures. Michael Shavlik and Michael Harms wrote the manuscript.

Abstract

As an evolving protein traverses sequence space, the "neighborhood" of accessible protein functions changes, thus shaping the evolutionary potential of the protein. Understanding how mutations alter the accessibility of new functions is important, both for understanding natural evolution and for engineering new protein function. To better understand how the functional neighborhoods of proteins change as they evolve, we studied the evolution of GFP-like proteins from the coral Montastrea cavernosa. Previous work showed the ancestral protein was green, but then acquired a photoconvertible green-red phenotype through substitutions scattered throughout the protein structure[51]. We combined random mutagenesis, flow cytometry, and high-throughput

sequencing to measure the frequency of different colors—green, orange, and red—up to six mutations away from a central genotype. We found that the neighborhood of the constitutively green ancestor contains substantially more functional proteins than that of a photoconvertible derived protein. Despite being less mutationally robust, the neighborhood surrounding the evolved protein showed greater phenotypic diversity than the ancestral protein. We also found we could dramatically alter neighborhood robustness by introducing individual historical substitutions, revealing that a single mutation can significantly change the local neighborhood of phenotypes accessible to an evolving protein. Preliminary biophysical characterization through HDX-MS suggests that the relative stability of local structural features near the chromophore, not the overall stability of the protein, modulates the landscape of these evolutionary neighborhoods. However, these results are preliminary and require further testing to be further supported. Subtle, local changes to protein structure can thus potentially profoundly alter evolutionary outcomes.

INTRODUCTION

Proteins evolve new functions through the accumulation of mutations. Their ability to successfully acquire mutations depends on the accessibility of functional genotypes nearby in sequence space [57–60]. Knowledge of the functional organization of relevant regions of sequence space, as well as the mechanisms that give rise to those properties, is critical to understanding the evolutionary process.

While there is no shortage of studies on functional characterization of protein sequence space [61–72], little work has been done to identify how the functional distribution of protein sequence space changes across evolutionary time for naturally evolved proteins[73]. Characterizing

small volumes of local sequence space ("neighborhoods") along a naturally evolved protein's evolutionary trajectory can reveal insights to how the accessibility of functions changed over time. Additionally, these types of observations could inform future predictive models for molecular evolution and have implications for protein engineering work to develop more robust protein structures [74].

Further, by drilling into the mechanisms that caused these changes, we can begin to generalize our results from a single protein to broader classes of proteins. Previous studies have shown that robustness can be controlled by global stability[75] and dynamics can improve evolvability [53,76–82]. Specifically investigating the interplay of these mechanistic relationships with natural evolution can reveal the biophysical features that underpin the accessibility of protein sequence space and thus the potential to develop new functional properties.

Fluorescent proteins are exceptional model proteins to use in mutational studies exploring sequence space. It is easy to detect changes or disruptions to the native color phenotype [65,83–89],. Additionally, since fluorescent proteins autocatalytically synthesize their chromophore functional group to achieve efficient fluorescence, we can potentially apply useful insights from the evolution of fluorescent proteins to understand the evolution of enzymes more generally [90].

In this study, we use a naturally evolved red photoconvertible fluorescent protein from the great star coral *Montastraea cavernosa*. The structure of this protein is consistent with most fluorescent proteins observed in nature, consisting of an 11-strand beta barrel enclosing a tripeptide chromophore essential in determining the color of fluorescence [91]. The protein forms its chromophore from three amino acids (H-Y-G). When expressed, this protein is green; upon exposure to UV light, it becomes red. The Matz group previously characterized the evolution of photoconversion using ancestral sequence reconstruction (ASR). They found that the modern

protein evolved from a constitutively green ancestral protein [51]. This evolutionary transition involved 37 substitutions, 12 of which are sufficient to recapitulate the evolution of photoconversion.

The historical substitutions exhibit intense epistasis. The most notable example is Q62H. This mutation converts the chromophore sequence from Q-Y-G to H-Y-G. The H-Y-G chromophore sequence is characteristic to all photoconvertible green-red fluorescent proteins [76,92]. When introduced alone, Q62H does not make the protein photoconvertible. Additional substitutions scattered throughout the protein structure, however, allow the Q62H substitution to confer photoconversion. Thus, in this transition, an appropriate genetic background is required for a few key functional substitutions to unlock green-to-red photoconversion.

This evolutionary transition allows us to pose several questions: How does protein mutational robustness and evolvability change over time? How does a single amino acid substitution alter the composition of the local sequence space, and how does the genetic background change that substitution's effects? What protein features determine the distribution of functional sequence space along an evolutionary trajectory?

To address these questions, we combined random mutagenesis and flow cytometry to characterize the local sequence space neighborhoods for the ancestral protein (ancGFP), the evolved photoconvertible protein with 15 substitutions from the historical trajectory (anc15), as well as several point mutants in both genetic backgrounds and evolutionarily unrelated fluorescent proteins. For each of these proteins, we assessed both the mutational robustness and phenotypic evolvability indicated by their local sequence space neighborhoods. We found that local neighborhoods in GFP-like protein sequence space change dramatically in their mutational robustness over evolutionary time, with evolved anc15 being much less robust than ancestral

ancGFP. Additionally, we found that a single amino acid substitution can significantly change the phenotypic diversity and mutational robustness of local neighborhoods. While global protein stability did not seem to directly control the neighborhood robustness of these proteins, we identified several features that may contribute to the composition of these neighborhoods. Kinetics of the photoconversion mechanism may present an evolutionary tradeoff between robustness and function in this protein's history, and preliminary data that has yet to fully be analyzed suggests stability of a few key local structures may tune mutational robustness. These results underscore the importance of studying small volumes of sequence space over real evolutionary time to bolster our understanding of how proteins navigate the evolutionary process.

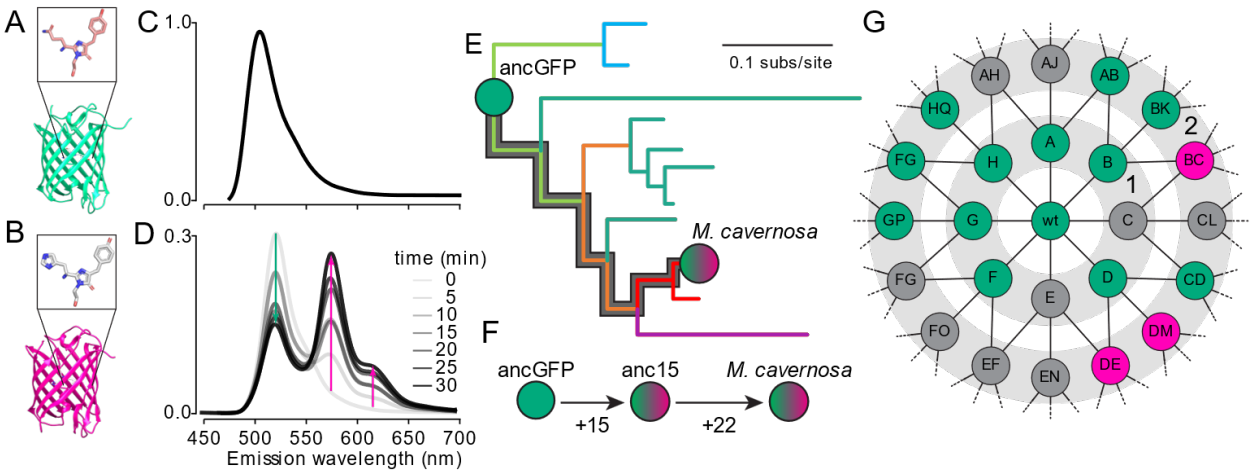

Fig 2.1. Characterization of ancGFP and anc15. A-B) Protein structure of ancGFP (A, PDB: 4DXI [53]) and anc15 (B, PDB: 4DXN [91]) with a close-up of the chromophore structure. C-D) Emission spectra of cells expressing ancGFP (C) and anc15 (D), excited at 380 nm. The anc15 plot shows spectra measured at 5-minute intervals upon exposure to 385-395 nm UV light. Green and pink arrows indicate peak intensity changes. E) Evolutionary tree used for original ancestral sequence reconstruction. Modern protein sequences were sampled from Faviina corals. The gray outline traces the evolutionary path between ancGFP (green circle) and its extant photoconvertible protein (green/pink circle). F) Number of substitutions between ancGFP, anc15, and the extant photoconvertible protein. G) Cartoon representation of a hypothetical local

evolutionary neighborhood around a fluorescent protein. Nodes indicate genotypes. Text indicates single mutations (e.g., “A”, “F”, etc.) and double mutations (e.g. “AB”, “FO”, etc.) Edges connect genotypes that differ by a single mutation. Node colors correspond to fluorescence color, with gray representing no detectable fluorescence.

RESULTS

AncGFP and Anc15 neighborhoods show different mutational robustness

We first estimated the local color neighborhood of ancestral GFP (ancGFP). This protein has been characterized previously, both functionally [92] and structurally [53,93]. AncGFP exhibits bright green fluorescence (Fig 1C). To estimate the color neighborhood, we characterized the fluorescence of bacteria transformed with nine random mutagenesis libraries. The pipeline is shown schematically in Fig 2. First, we generated libraries with nine different mutation rates. We then used high-throughput sequencing to estimate the frequency of each mutant class (no mutations, one mutation, two mutation, etc.) in each library (Fig 2A-C). We transformed bacteria with the libraries, aiming for ~10,000 genotypes per library. We then used flow cytometry to measure the emission color of 100,000 individual bacteria per library (900,000 total observations) (Fig 2D, 2E). To measure the number of non-fluorescent library members, we expressed the ancGFP mutants as a fusion construct with a protein that fluoresces in the infrared (miRFP703) [94]. We classified proteins emitting in the infrared but not visible wavelengths as non-fluorescent. Finally, we inferred the distribution of colors in the local neighborhood by combining the measured frequencies of the mutant classes and colors in each library (Fig 2F). We used a linear mixture model to infer this distribution of colors in the neighborhood using these frequencies of mutant classes and observed colors via cytometry. The model returns the expected proportion of proteins with a certain color given the number of mutations, ranging from 0 to 6+. (See methods for details).

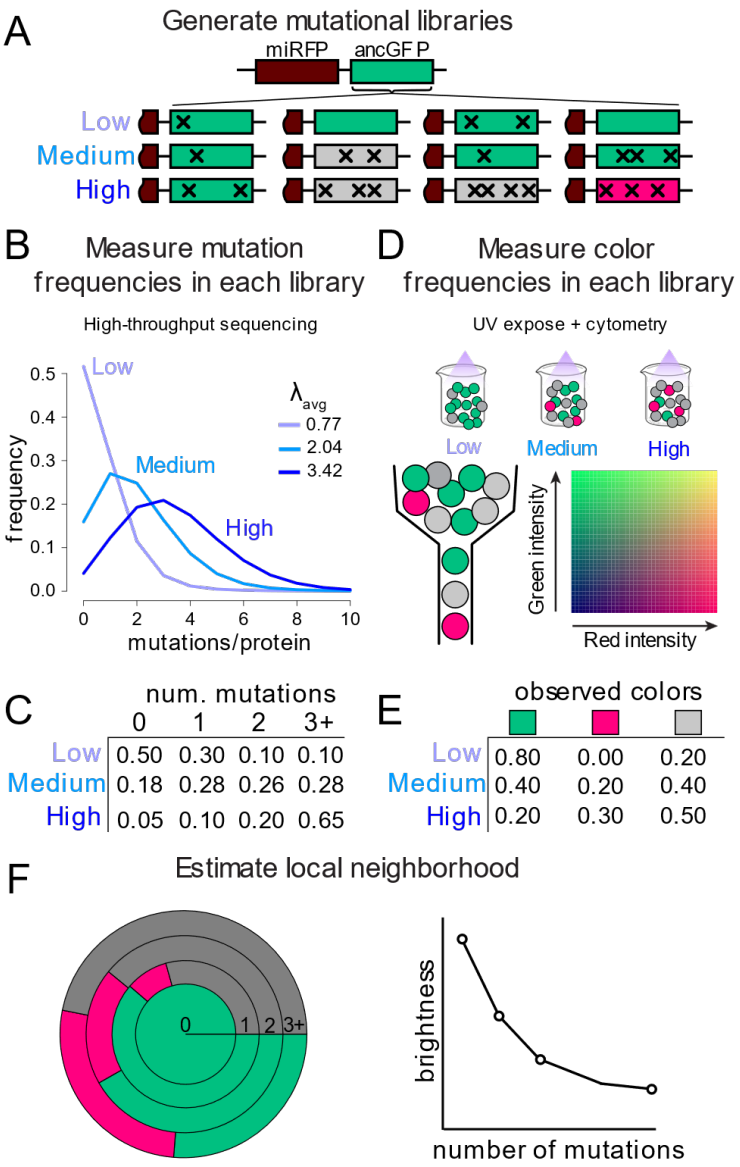

Fig 2.2. Characterizing a protein's local neighborhood. For simplicity, we illustrate this process using three libraries; in our experiments we used nine libraries with different mutation rates. A) We expressed ancGFP (green) as a fusion construct with miRFP703 (maroon). We generated ancGFP libraries with increasing mutation rates (indicated by "×" icons), leaving miRFP703 untouched. The color of each mutant is indicated, with gray representing non-fluorescent proteins. B) We measured the mutation rate of each library using high-throughput sequencing. Darker blue lines indicate higher average mutation rate. C) Table of mutant class frequencies determined by the sequencing calibration for each library shown in panel B. D) We transformed each library into *E.coli* and measured its color using flow cytometry. Color grid represents 1600 different red/green intensity bins in which a mutant protein could be categorized. E) Cytometry color frequencies seen in each library (out of 100,000 observations/library). Note:

for simplicity we show three colors; in our analysis we used all 1,600 color bins above. F) W used a linear mixture model to infer the local neighborhood color (left) and brightness (right) distributions from the mutation frequencies and cytometry counts from panels C and E. Gray represents non-functional proteins. Each ring (left) indicates protein phenotypes with N number of mutations.

We found that the neighborhood of ancGFP was robust and uniform. Even after accumulating ≥6 mutations, 84% of the variants exhibited green fluorescence (outermost ring, Fig 3A). Further, the average brightness of clones is unchanged, even out to 6 mutations (Fig 3B). Because we measured the colors of ~10,000 clones per library, a phenotype would only appear in this analysis if it occurred at a frequency greater than 1/10,000 (see methods for details). We can therefore say that at the farthest region of the neighborhood characterized (≥ 6 mutations), 84% of the genotypes were green, 16% were non-functional, and <0.1% were any other color. This reveals that the ancestral protein could access many sequences, but these contained minimal color variation. This lack of color diversity is in line with previous studies, which demonstrated that multiple historical mutations were necessary in this trajectory to induce any notable color changes from this ancestor [52,53,92].

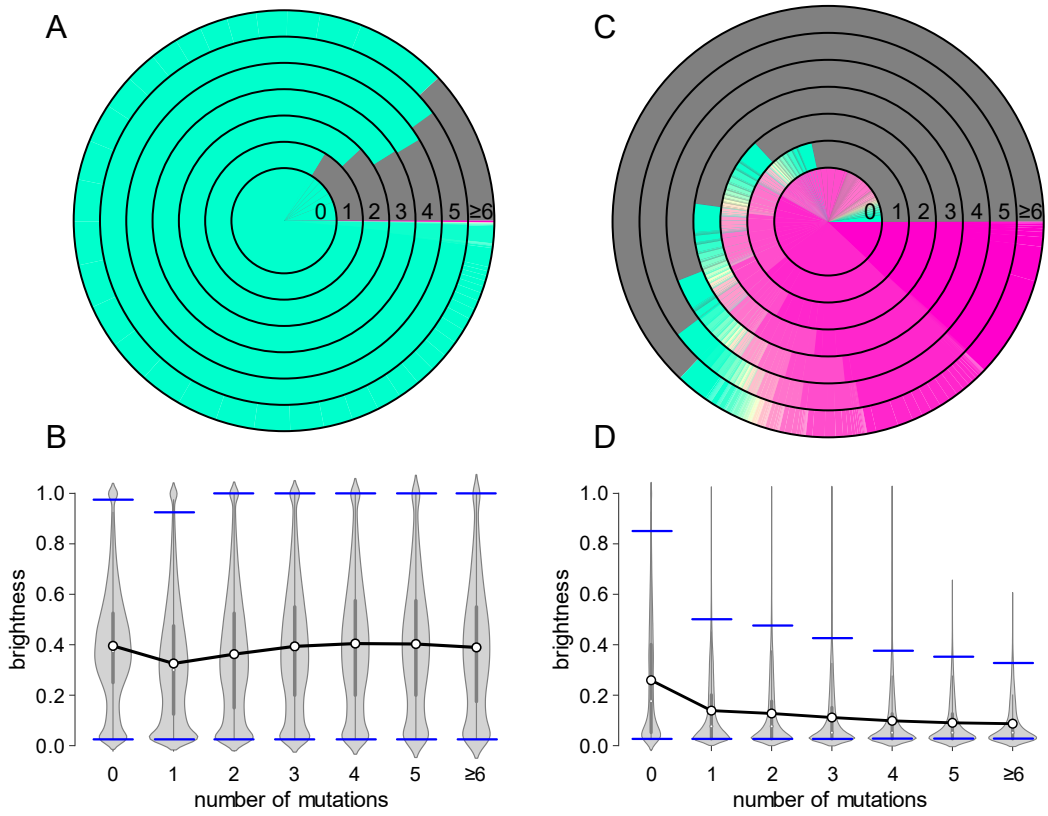

Fig 2.3. Local neighborhoods of ancGFP and anc15. A) Neighborhood of ancGFP. Rings indicate the fraction of proteins with N mutations (from 0 to ≥6) that have the indicated color. Gray regions indicate proteins that had no detectable fluorescence. B) Brightness distribution of proteins with N mutations shown as a violin plot. Proteins are normalized to the brightest 5% of the neighborhood. Mean brightness for proteins with each number of mutations is indicated by the black line and white points. Blue lines indicate the bounds for the middle 95% of the brightness values observed for each mutation number. Thicker regions of the violin indicate higher density of proteins at that relative brightness. The black bars inside of each violin show the interquartile range of the distribution. Thinner gray bars within each violin stretch to the minimum and maximum values that are not outliers. This analysis only includes proteins that fluoresce above background (i.e., non-gray proteins from panel A). C) Neighborhood of anc15 after 30 minutes of UV exposure for photoconversion. D. Brightness distributions for anc15 with means and 95% bounds indicated the same as ancGFP.

We next asked how the local neighborhood changed as the protein evolved. We introduced 15 historical substitutions into ancGFP (Fig 1F). This "anc15" ancestor is similar to the so-called Least-Evolved Ancestor (LEA) previously characterized [51,52]. The LEA protein consists of ancGFP with twelve historical substitutions; it is functionally indistinguishable from a modern photoconvertible red protein. Anc15 contains the 12 substitutions used for the LEA protein, with an additional three historical substitutions (E26V, D74H, and V214E). These substitutions occurred between ancGFP and the modern *M. cavernosa* photoconvertible protein[51,52]. (These specific substitutions were added to the LEA for a different project in the lab. They do not significantly change the phenotype from the LEA with 12 historical substitutions). Like the LEA, anc15 is photoconvertible and much dimmer than ancGFP: it behaves indistinguishably from a modern photoconvertible red protein (Fig 1C, 1D).

We characterized the sequence neighborhood of anc15 and compared it to that of ancGFP. We found that anc15 was much less robust than ancGFP (Fig 3C, 3D). Only 36% of the variants with ≥6 mutations exhibited detectable fluorescence. In addition, the proteins that were fluorescent were, on average, much dimmer across the anc15 neighborhood compared to ancGFP (Fig 3B vs. 3D). Additionally, the anc15 neighborhood had a higher diversity of colors than ancGFP, exhibiting red, orange, and green phenotypes. The presence of green and orange clones is unsurprising, as anc15 is green prior to photoconversion. Failed or incomplete photoconversion would lead to a green or orange phenotype.

Validation of high-throughput results

To validate our high-throughput results, we measured the effects of 5 individual point mutants on the fluorescence of ancGFP and anc15. We expressed each mutant protein in bacteria and then measured their fluorescence intensity in a plate reader. The mutations in question were

I56N, Q/H62R, N65K, P72S, and K80E. The sites were in the inner core of the beta-barrel, ranging from the bottom of the chromophore-bearing pocket through the top of the protein (Fig 4A). Given their locations adjacent to the chromophore, these mutations were likely to be influential in maintaining a proper fluorescence phenotype. These mutations sample a variety of shifts in amino acid properties, including changes in hydrophobicity, side chain length, and charge states.

We found that these mutations had dramatically different effects on anc15 versus ancGFP. In anc15, every mutation disrupted fluorescence, with both the green and red wavelengths falling near or below cellular autofluorescence (Fig 4B). This result is consistent with the mutationally sensitive neighborhood observed for anc15 in Fig 3C and 3D. In contrast, mutations in ancGFP displayed a range of effects. Consistent with the neighborhood of ancGFP, some mutations dimmed (I56N) or completely ablated fluorescence (Q62R and N65K), while others maintained a roughly equivalent or higher brightness (P72S and K80E) (Fig 4B).

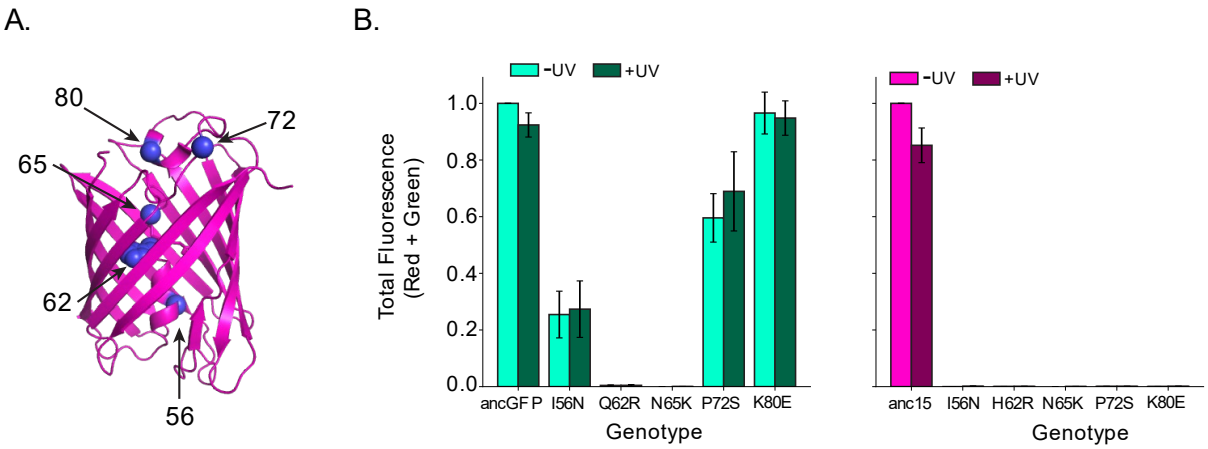

Fig 2.4. Experimental characterization of selected library mutants in ancGFP and anc15 backgrounds. A) Locations of five mutations, shown on the structure of LEA (PDB: 4DXN [91]). Spheres indicate alpha carbons, except for position 62 where the whole amino acid is shown as spheres. B) Fluorescence intensity of ancGFP (left) and anc15 (right) mutants relative to the associated wild type at combined 510 nm and 585 nm emission wavelengths (roughly corresponding to the characteristic green and red emission spectrum peaks in anc15, figure 1D). Intensities are summed from independent excitation measurements of 405, 488, and 561 nm. Intensities are shown either with no prior UV exposure (-UV) or with 30 minutes of prior UV exposure (+UV). Bar is mean intensity across three biological replicates; error bars represent standard error of the mean. Darker bars are measured intensities after 30 minutes of UV exposure. Fluorescence is shown in arbitrary fluorescence units above cellular autofluorescence and normalized for differences in expression level by using the fluorescence intensity of a fused and unmutated miRFP703 at 703 nm emission. Green and red emissions of all mutants are shown in supplemental figure 6.

Single substitutions can alter the neighborhood

The dramatic difference in the neighborhoods of ancGFP and anc15 reveals that local volumes of sequence space can change quite dramatically over evolutionary time. While this finding is significant, the transition from ancGFP to anc15 involves 15 substitutions—a substantial jump in sequence space. Because of this jump, it is difficult to determine whether these changes in neighborhoods are a result of many small-effect size mutations or few large-effect size mutations. To ask this question, we made single-site substitutions to both ancGFP and anc15 and measured the resulting changes in sequence neighborhoods.

To probe the effects of single mutations, we characterized the effects of two historical mutations on the local neighborhoods of ancGFP and anc15 (Fig 5A). The first mutation was Q62H (Fig 5B). H62 is directly incorporated into the chromophore during photoconversion and is thus critical for photoconverted red color of the protein. The derived histidine is conserved across the derived photoconvertible fluorescent proteins [83,90,92]. While this histidine has been shown previously to be necessary for photoconversion, it is not sufficient for photoconversion when introduced alone into the ancGFP background [51,53]. In addition to the lack of color changes associated with the histidine substitution alone, this substitution has not been found to induce large structural changes to the ancGFP structure [95]. Together, previous literature suggests a small effect size, if any, of the single Q62H substitution in re-shaping the neighborhood of ancGFP.

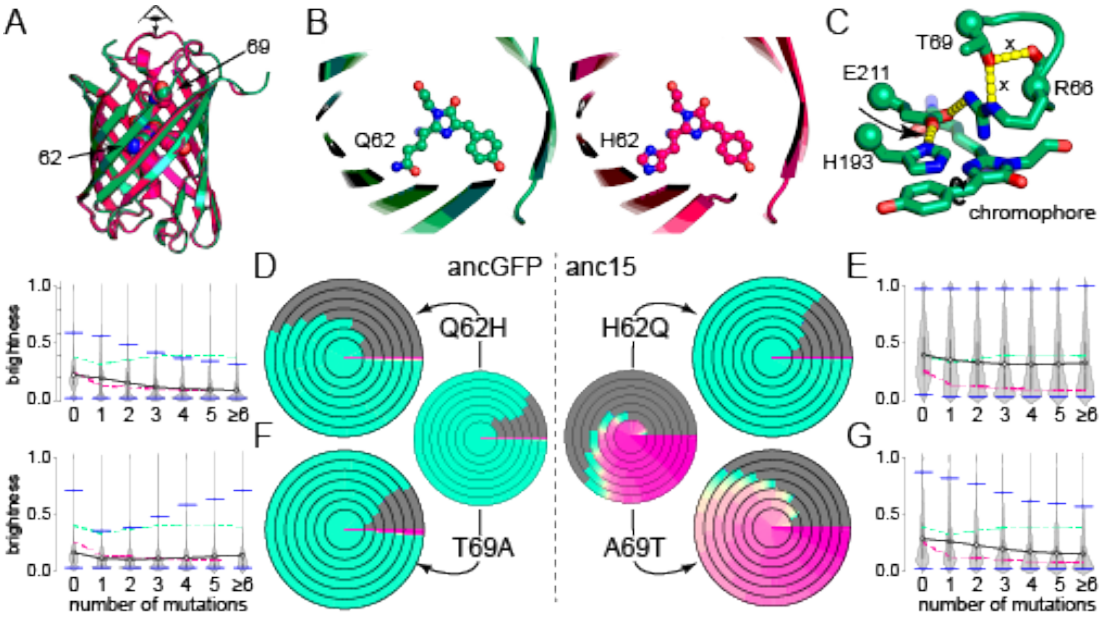

Fig 2.5. Point mutants of ancGFP and anc15 and their local neighborhoods. A) Structures of ancGFP (green) and anc15 (pink) indicating positions 62 and 69 via arrows (PDB: 4DXI, 4DXN). Eye icon indicates the viewpoint shown in next panel. B) Top-down view of the chromophore in ancGFP and anc15 showing Q62 in ancGFP and H62 in anc15. C) Structure shows a hydrogen bond network (yellow dashes) that influences photoconversion rate. The black arrow indicates the hydrogen bond that must break during photoconversion. The × icons indicate the hydrogen bonds disrupted by the T69A substitution. D-G) Color neighborhood and brightness distributions of point mutants. D) Smaller ring diagram reproduces ancGFP neighborhood from Fig 3A; larger ring diagram shows ancGFP/Q62H. Violin plot shows brightness of genotypes within each ring. All graphical elements match those shown in Fig 3 A and B. Green and red dashed lines on brightness graphs represent the averages of ancGFP and anc15, respectively (Fig 3B,D). E) Results for anc15/H62Q. F) Results for ancGFP/T69A. G) Results for anc16/A69T.

We found that mutating Q62H in ancGFP resulted in a less robust neighborhood with little to no impact on color diversity (Fig 5D). While 84% of the ancGFP neighborhood exhibited fluorescence, only 57% of the ancGFP/Q62H neighborhood did. Further, the brightness of the proteins that did fluoresce was substantially dimmer, being much more similar to anc15 than ancGFP (Fig 5D). We next tested how the H62Q reversion in anc15 altered its neighborhood. This mutation dramatically increased the robustness of the local neighborhood and eliminated its color diversity (Fig 5E). It also increased the brightness of the proteins that exhibited fluorescence (Fig 5E). Overall, the anc15/H26Q neighborhood was indistinguishable from the ancGFP neighborhood both in terms of colors observed and brightness (Fig 5E). The lack of color diversity is unsurprising, as histidine is critical for formation of the red chromophore; however, the massive change in robustness due to a single point mutation was unexpected given the documented lack of functional and structural changes from the single Q62H substitution.

The Q62H mutation, by itself, thus acts as a toggle. In the Q state, the neighborhood is more robust and brighter, but loses all color diversity. In the H state, the neighborhood is less robust and dimmer. H interacts epistatically with the background. In ancGFP, it has no effect on color diversity; in the anc15 background, it confers photoconversion and thus red and orange phenotypes.

We next asked how a different historical substitution altered the sequence neighborhood. The T69A mutation dramatically increases the rate of photoconversion by a known mechanism [53,76,90]. Chromophore maturation requires the rotation of its phenol ring, which in turn requires breaking the ion pair between His-193 and Glu-211 (Fig 5C). The historical T69A mutation removes two hydrogen bonds to Arg-66 (indicated by "×" in Fig 5C), thus destabilizing the polar network involving His-193 and Glu-211. This, in turn, speeds up photoconversion.

We introduced the historical T69A mutation into the ancGFP background and measured its local neighborhood. We found no change in the fraction of fluorescent proteins and no appreciable difference in color diversity compared to ancGFP. There was, however, a large drop in the brightness of sequence neighborhood, with mutants in the ancGFP/T69A neighborhood resembling anc15. This is consistent with solvent quenching due to increased dynamics in the region "above" the chromophore in the beta barrel [96].

We next reverted A69T in the anc15 background. We saw an appreciable increase in robustness (61% fluorescent in the outer ring of A69T versus 36% in anc15), as well as retention of the diverse color phenotypes (Fig 5G). We also saw a moderate increase in brightness. It is important to note, however, that 3x the UV-exposure time was required to achieve high levels of photoconversion across the neighborhood when compared to anc15 (see next section).

Taken together, these results underscore the importance of the historical Q62H and A69T mutations on the mutational robustness of the evolving protein. Additionally, these results suggest that these specific residue changes can toggle the accessibility of immediate sequence space in this trajectory. These results show that a single site substitution involving function optimization can significantly impact the neighborhood, but the effect size of that substitution may depend heavily on genetic background.

Photoconversion does not affect neighborhood robustness

We wanted to test if the process of photoconversion itself was responsible for decreasing neighborhood robustness in these two proteins. One possibility was that ancGFP and anc15 had similar levels of robustness for the initial green color, but that mutations in anc15's neighborhood resulted in non-productive photoconversion and thus ablation of fluorescence instead of a normal green-to-red color shift. If non-productive photoconversion is pervasive and contributes to higher

mutational sensitivity and decreased robustness in the neighborhood, we would expect the fraction of non-fluorescent proteins to increase as a function of UV exposure (Fig 6A). Alternatively, if the mechanism of photoconversion does not affect the neighborhood robustness, the fraction of non-fluorescent proteins should remain relatively constant given UV exposure time (Fig 6B). We measured the neighborhood of anc15 in 5-minute intervals of UV exposure time and found no significant change in the fraction of non-fluorescent mutant proteins (Fig 6C). As expected, the rate of photoconversion decreased for the mutant proteins compared to anc15 with 0 mutations (Fig 6D). Similarly, anc15 A69T showed no significant change in the fraction of non-fluorescent proteins across 15-minute intervals of UV exposure time (Fig 6E). Mutated anc15 A69T also showed some decreased rates of photoconversion efficiency (Fig 6F). These results indicate that the mechanism of photoconversion does not affect neighborhood robustness and that differences observed between neighborhoods is likely due to some other protein feature(s).

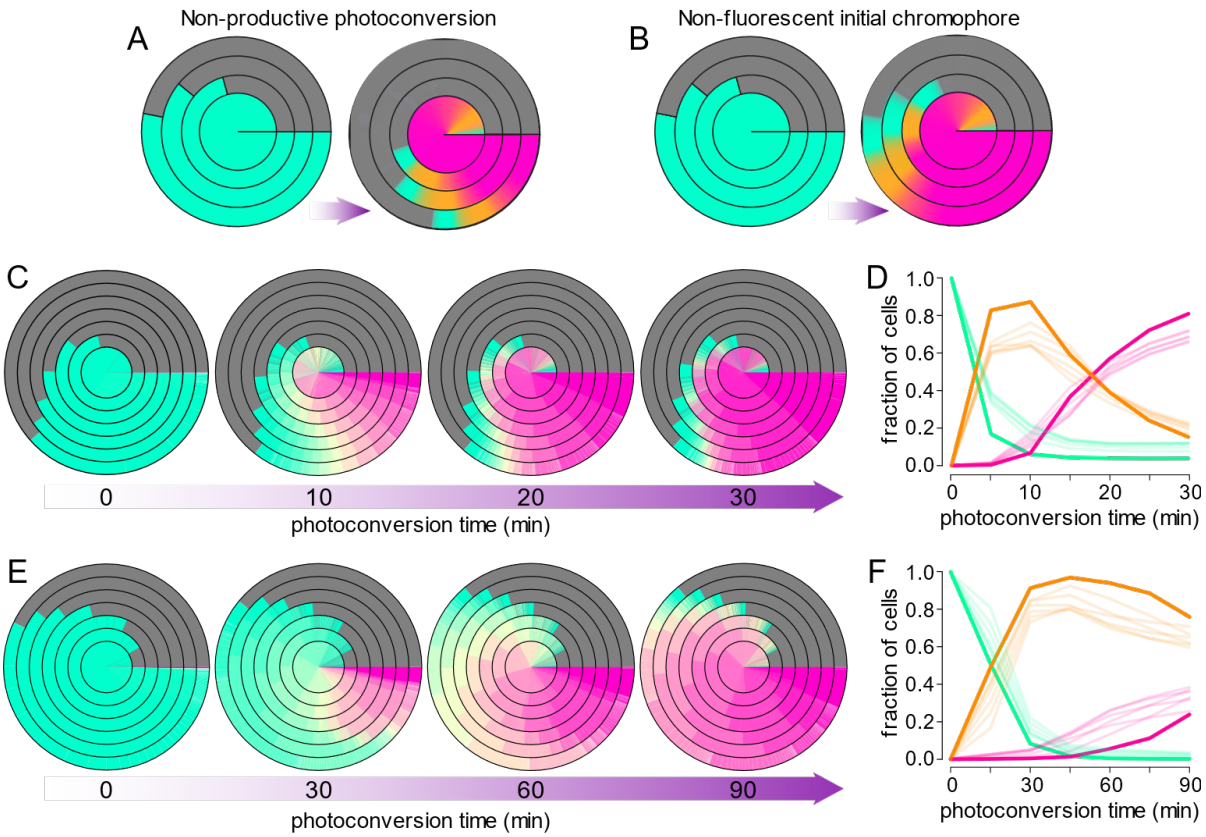

Fig 2.6. Neighborhoods of photoconvertible proteins over UV exposure time. A-B) Cartoons depicting the expected outcome if decreased mutational robustness arises from non-productive photoconversion (A) or some other mechanism (B). C) Neighborhood of anc15 across 30 minutes of UV exposure time. D) The fraction of green, orange, and red proteins in each ring of anc15's neighborhood over photoconversion time. The bold line tracks the center ring of non-mutant proteins while the thinner lines follow each mutant ring. E-F) Results for anc15/A69T, as in panel C-D. Note that this required 90 minutes of UV exposure time.

Unrelated RFPs are not mutationally robust

We next wanted to further probe why local neighborhoods differ from one another in their robustness. One possible explanation for why ancGFP is more robust than anc15 is because anc15 is a statistically inferred photoconvertible intermediate protein and not an extant protein. Since anc15 is statistically inferred, it could possess a pathological, non-historical combination of mutations that artificially compromises robustness. To test this possibility, we generated the local neighborhood of an extant photoconvertible fluorescent protein, meleRFP, derived from *Mycedium elephantotus* [64]. Since meleRFP is unrelated to *M. Cavernosa*, its neighborhood additionally reveals whether low mutational robustness is specific to photoconvertible proteins in the historical trajectory that includes anc15. We found that the local neighborhood of meleRFP is very similar to anc15 in both robustness and phenotypic diversity, with only ~20% of proteins with 6+ mutations fluorescing either green, orange, or red above background fluorescence (Fig S2).

Another possibility is that having the ability to photoconvert is itself causing anc15 to be less robust. Since this is a feature that we have observed with both anc15 and meleRFP, we decided to estimate the local neighborhood of a constitutively red fluorescent and engineered protein mCherry [88]. We found that despite having a different mechanism of chromophore maturation to achieve red fluorescence, mCherry and anc15 have similar mutational robustness (Fig S2). These results indicate that changes in neighborhood composition are likely not due to anc15 being an evolutionary intermediate and that a lack of mutational robustness is not a feature exclusive to photoconvertible fluorescent proteins.

Protein stability does not explain neighborhood robustness

We next sought a biophysical explanation for the huge changes in robustness conferred by mutations at sites 62 and 69. Previous work has shown that protein stability can modulate the evolvability of proteins [75,97–99]. Higher stability allows a protein to tolerate destabilizing mutations, and thus could increase the functional volume of the sequence neighborhood. To determine the extent to which stability controls robustness for these proteins, we measured the stability of ancGFP, anc15, GFPQ62H, GFP/T69A, anc15/H62Q, and anc15/A69T using guanidinium hydrochloride. We found that the denaturation curves could be fit with an apparent two-state folding model (Fig 7A). We saw little correlation in the denaturation midpoint and the robustness of the map. (Fig. 7B). We repeated the stability measurement using temperature denaturation and again found no correlation between melting temperature (irreversible) and mutational robustness (Fig S3). This suggests other factors besides global stability control the robustness of these proteins to mutation.

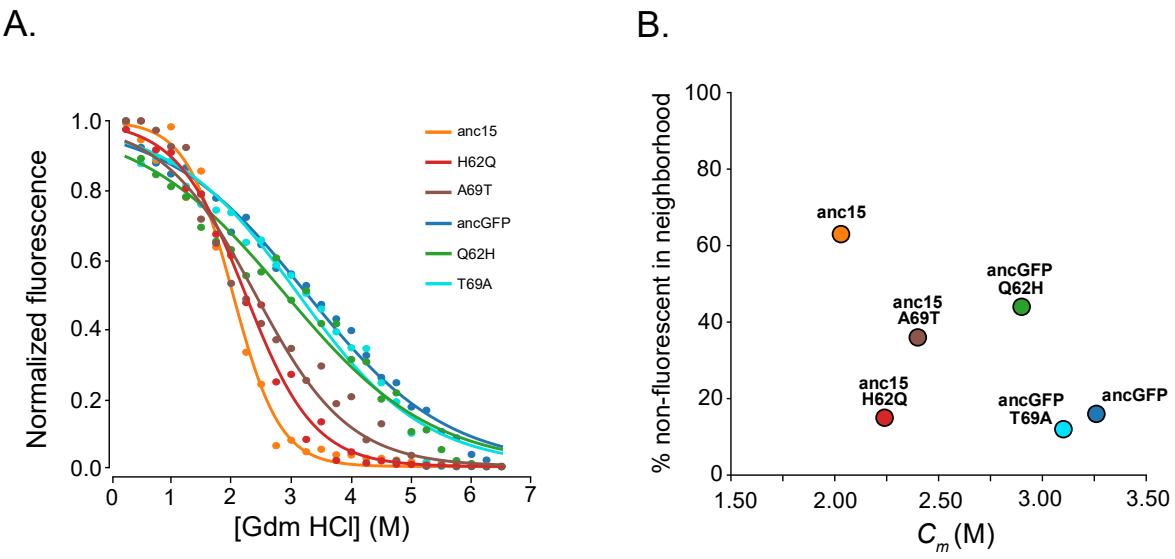

Fig 2.7. Protein stability does not correlate with mutational robustness. A) Normalized fluorescence measurements (emission at 508/518 depending on genetic background) across 0.25M intervals of GdmHcl. After mixing, samples equilibrated for 2 weeks to assure equilibrium. We fit a 2-state unfolding model to the unfolding curves to extract C_M. Fit parameters are given in table S1. B) Extracted C_m parameters from panel A for each protein genotype versus the fraction of non-fluorescent proteins in the ≥6 mutation ring from each neighborhood.

DISCUSSION

Here, we introduced a new method of to rapidly quantify changes in local sequence neighborhoods for fluorescent proteins. This method is useful for high-throughput characterization of thousands of mutant fluorescent proteins using RGB intensities and can, from mutant library generation to a neighborhood diagram, be completed in as little as a week. Through applying this method to a naturally evolved GFP-like protein, we demonstrate that changes in mutational robustness can be profound from one local neighborhood in sequence space to another. While we have shown that a local neighborhood can change dramatically between relatively distant regions of sequence space (i.e. ancGFP vs anc15), we also found that a

single point mutation can significantly re-shape the neighborhood accessibility in both genetic backgrounds.

While the underlying mechanism or structural feature responsible for modulating local neighborhood robustness is unclear, we found little supporting evidence for a few possible features in this study. First, global structural stability does not seem to directly correlate with neighborhood robustness. While a single amino acid change at position 62 or 69 can shift both the stability and robustness of ancGFP or anc15, these two features appear to be decoupled (Fig 7B). Second, the process of photoconversion itself does not drive the divergence in robustness between ancGFP and anc15 (Fig 6). Despite this, the apparent tradeoff between photoconversion efficiency and neighborhood robustness observed with anc15A69T is interesting and may suggest a potential role for dynamics and protein flexibility near the chromophore as one driver for shaping local neighborhoods.

Stability on the local scale of important structural features near the chromophore bearing pocket may also be tuning the mutational robustness of these proteins. We are currently running Hydrogen-Deuterium Exchange- Mass Spectrometry to measure the stabilities of local peptides in each of the analyzed proteins in this study. We aim to include this piece of data in the final iteration of this paper.

This study used a novel high-throughput pipeline for generating local neighborhoods for a naturally evolved protein at multiple points along a historical trajectory. Though the mechanism responsible for determining the accessibility of the neighborhoods remains unclear, we show evidence against a few possibilities and lay the groundwork for potential avenues in exploring the mechanisms that do directly shift local regions of sequence space. Future studies should focus on further elucidating the effects of local structural stability or dynamics on

neighborhood robustness and assessing the generality of those effects to broader classes of enzymes and their evolution.

MATERIALS AND METHODS

Linear modeling for neighborhood inference

We used a linear model to infer the color neighborhood for each genotype. Imagine we wish to measure the distribution of green genotypes in a neighborhood that consists of genotypes with zero, one, or two mutations. The probability that a randomly selected clone is green is given by:

$$P(G) = P(G|n_0)P(n_0) + P(G|n_1)P(n_1) + P(G|n_2)P(n_2)$$

where $P(n_i)$ is the probability that a clone has i mutations and $P(G|n_i)$ is the probability that a clone with i mutations is green. Our goal is to estimate the values of $P(G|n_0)$, $P(G|n_1)$, and $P(G|n_2)$, thus characterizing the green neighborhood from 0 to 2 mutations away from the initial genotype.

To estimate these values, we can construct libraries that vary in their proportions of genotypes with 0, 1, or 2 mutations. For example, we might make three libraries. The first library has only clones with zero mutations and no clones with one or two mutations. The second library has an even mixture of genotypes from each class: 1/3 of the clones have no mutations; 1/3 have one mutation; and 1/3 have two mutations. The third library has a different proportion of genotypes from each class: 1/10 of the clones have no mutations, 4/10 have one mutation, and 5/10 have two mutations. If we substitute in the proportions of each genotype class within each library into $P(n_i)$, we obtain the following three equations for the probability we see a specific color in each of the three libraries:

$$P(G)_{lib1} = P(G|n_0) \times 1 + P(G|n_1) \times 0 + P(G|n_2) \times 0$$

$$P(G)_{lib2} = P(G|n_0) \times 1/3 + P(G|n_1) \times 1/3 + P(G|n_2) \times 1/3$$

$$P(G)_{lib3} = P(G|n_0) \times 1/10 + P(G|n_1) \times 4/10 + P(G|n_2) \times 5/10$$

If we can experimentally measure $P(G)_{lib1}$, $P(G)_{lib2}$, and $P(G)_{lib3}$, we now have three equations with three unknowns and can thus solve for $P(G|n_0)$, $P(G|n_1)$, and $P(G|n_2)$. This can be generalized for any color, up to *L* mutations as:

$$P(color) = P(color|n_0)P(n_0) + P(color|n_1)P(n_1) + \cdots P(color|n_L)P(n_L).$$

We generated libraries with different values for $P(n_i)$ by error-prone PCR. This generates a Poisson-distributed set of mutations controlled by the mutation rate λ. By varying the mutation rate between libraries, we can build libraries with known values for $P(n_i)$. Our final model is thus:

$$P(color|\lambda) = \sum_{i=0}^{i<L} P(color|n_i)P(n_i|\lambda)$$

where $P(color|\lambda)$ is the probability of seeing a specific color given a library mutation rate λ and $P(n_i|\lambda)$ is the probability of observing a clone with i mutations given the library mutation rate λ.

For each neighborhood, we generated 8 libraries with mutation rates between an average of 0.77 mutations per gene and 3.42 mutations per gene. Having 8 libraries allowed us to write 8 equations with 8 unknowns, and thus measure the sequence space out to 7+ mutations.

We measured $P(color|\lambda)$ for each library by flow cytometry (details below). We defined color as using the values of the red, green, and blue cytometer channels. We broke each channel into 40 evenly spaced bins. We could thus resolve the 40 × 40 × 40 = 64,000 different colors in each experiment. We experimentally estimated $P(n_i|\lambda)$ for our libraries using high-throughput sequencing (see below).

Construct design

We expressed all ancestral GFP-like proteins as a fusion, with an engineered N-terminal bacterial phytochrome miRFP703 [94] connected to a C-terminal GFP-like protein by a 40 amino acid long rigid linker that predominantly consisted of alanine residues. We used this linker to reduce Forster Resonance Energy Transfer (FRET) efficiency between the two proteins [100]. This fusion construct design allowed us to detect the presence of non-fluorescent GFP-like proteins. The miRFP703 phytochrome fluoresces with a peak intensity of 703 nm while the GFP-like proteins fluoresce with peak intensities between 450 and 650 nm. By measuring fluorescence on channels across these wavelengths, we can not only detect shifts in blue, green, and red wavelengths from our GFP-like protein, but also when the only fluorescent molecule in the construct is the miRFP703.

Library plasmid and cell preparation

We used error-prone PCR to introduce random mutations into the GFP-like protein and re-synthesized the vector backbone using the Agilent GeneMorph EZ clone domain mutagenesis kit. We mutated residues 31-227 (Fig S1). Once the mutant library was made, we transformed XL10-Gold ultracompetent cells via heat shock at 42°C. Around 10,000 colonies were then collected and aliquoted to a cryogenic tube in a 1:1 mix with 50% glycerol and stored in a -80°C freezer.

Cytometry sample preparation

For each fluorescent protein sample, we grew overnight cultures from glycerol stock cells at 37 °C in LB broth with dual selection antibiotics of chloramphenicol and ampicillin. We extracted 700 µL of the overnight culture and transferred it to 25 mL of fresh LB broth + antibiotics. We then grew these cells at 37 °C until an optical density between 0.6-0.8 was reached. To induce expression, we added 25 µL of 1M IPTG and then allowed cells to express for 6 hours at 37 °C. Following expression, we pelleted cells by centrifugation, discarded the supernatant, and stored the pellets in a -20 °C freezer overnight. The next day, we washed cells with 1X PBS and centrifuged at 5,000 RPM for 5 minutes at 4 °C. Following centrifugation, we discarded the PBS, and fresh 1X PBS was added to resuspend the cells. If photoactivation was required, we exposed the cells to a UV flashlight for 30 minutes. We diluted each sample 1:10 in 1X PBS in a black, clear bottom 96-well plate for triplicate OD measurement with an i3x SpectraMax plate reader. Finally, we diluted the cells to an OD of 0.4 and placed them on ice.

Flow cytometer specifications

We ran our cells through a SonySH800 conventional flow cytometer available through the Genomics and Cell Characterization Core facility at the University of Oregon. To maximize the probability of exciting a fluorescent protein, we used 405, 488, 561, and 638 nm lasers in all experiments and acquired the fluorescence emission using fluorescence channels 1-5 (in nm: 425-475, 500-520, 565-600, 640-680, 690-750). To reduce signal overlap between channels 2 and 3, we exchanged their bandpass filters and reduced the bandpass width to 510 (+- 10 nm) and 585 (+- 15 nm), respectively.

Cytometry data calibration

To reduce day-to-day variability and discrepancies in laser gain settings across experiments, we calibrated every experiment to standardized fluorescence units using Spherotech 8-peak rainbow calibration beads (RCP 30-5A). We diluted these beads by adding 2-3 drops of the solution into 1 mL of 1X PBS. We then used this bead data to standardize our experiment using the open-source software package FlowCal in Python [101]. We executed each calibration step according to the sample workflow documented on the FlowCal ReadtheDocs (https://flowcal.readthedocs.io/en/latest/python_tutorial/index.html).

Library mutation rate calibration

We measured the frequency of clones with 0, 1, 2 … 6+ mutations in each error-prone PCR using Illumina NovaSeq 6000 paired-end 150bp sequencing. We performed this sequencing on each library in the ancGFP genotype background. We used a standard Phusion polymerase PCR amplification on each sample and then purified the resulting amplicons through GeneJet gel purification. Following purification, we ligated Illumina TruSeq adapters and finalized libraries for high-throughput sequencing following the Roche Kapa Hyperprep protocol. For each library, we calculated the average per-gene mutation rate. From there, we compared the theoretical Poisson mutation frequency distribution to the observed frequency distributions with an average R^2 of 0.992. These observed mutation frequencies were used in the linear model for inferring the local neighborhoods (Fig S4).

Measuring protein stability

We measured protein stability by first fully denaturing each protein in 7M guanidinium hydrochloride (GdmHCl). We left each protein to sit for 1 week in this high concentration solution before transferring to lower concentrations of Gdm. We used black 1.5 mL microcentrifuge tubes to prevent light exposure photoconverting any photoconvertible competent proteins during the chemical denaturation. After sitting for one week, we created solutions across a range of guanidinium hydrochloride concentrations [GdmHCl]. These concentrations of GdmHCl ranged between 0M to 7M in intervals of 0.25M as measured by a refractometer[102]. We added each protein to each solution such that the final concentration in each solution was 10 μM. We did this step so proteins could re-fold and equilibrate to a balance of unfolded and folded states. After 1 additional week of equilibration and re-folding, we measured fluorescence emission at either 508/518 nm on the i3x SpectraMax plate reader depending on which protein was being analyzed. We fit a 2-state folding model on these fluorescence emission results for each protein and extracted the CM (concentration at which fluorescence was half of the maximum observed) as a descriptor of protein stability.

Bridge to chapter III

In chapter II, I demonstrated the utility of applying high-throughput mutant library generation with experimental characterization to a naturally evolved protein system using GFP-like proteins. Not only does mutational robustness change dramatically over short periods of evolutionary time for the protein examined, but the biophysics that underlie this change are complex and potentially involve local structural stability or tradeoffs between kinetics and function. This study reveals a powerful method that can be used moving forward to further use the lens of biophysics in exploring protein sequence space. While this technique from chapter II is useful in exploring swaths of sequence space around a historical trajectory, we can also use the historical trajectory itself to gain insight into the biophysics guiding protein evolution. Ancestral Sequence Reconstruction, as mentioned prior, is an excellent technique to resurrect an ancient protein and regenerate the likely historical trajectory of the protein's evolutionary path. Using this technique coupled with a biophysical lens, we can examine the changes to a protein's energy landscape over time and investigate how much of a driving force this landscape is for evolution. In chapter III, I introduce this concept of the protein energy landscape, give examples of how we can tie this idea into ASR studies, and provide future directions of potential avenues for the field moving forward.

CHAPTER III

ANCESTRAL RECONSTRUCTION AND THE EVOLUTION OF PROTEIN ENERGY LANDSCAPES

Author contributions

Michael Harms, Lauren Chisholm, Kona Orlandi, Sophia Phillips, and Michael Shavlik conceptualized the study. Each author contributed writing and editing to the manuscript and text was finalized by Michael Harms. Figures were created and edited by each author and finalized by Michael Harms. Michael Harms administered the tasks for the project.

Abstract

A protein's sequence determines its conformational energy landscape. This, in turn, determines the protein's function. Understanding the evolution of new protein functions therefore requires understanding how mutations alter the protein energy landscape. Ancestral sequence reconstruction (ASR) has proven a valuable tool for tackling this problem. In ASR, one phylogenetically infers the sequences of ancient proteins, allowing characterization of their properties. When coupled to biophysical, biochemical, and functional characterization, ASR can reveal how historical mutations altered the energy landscape of ancient proteins, allowing the evolution of enzyme activity, altered conformations, binding specificity, oligomerization, and many other protein features. Here, we review how ASR studies have been used to dissect the evolution of energy landscapes. We also discuss ASR studies that reveal how energy landscapes

have shaped protein evolution. Finally, we propose that thinking about evolution from the perspective of an energy landscape can improve how we approach and interpret ASR studies.

Introduction

The sequence of a protein encodes a conformational energy landscape [103,104]. Some conformations are favored, others less so. The ensemble of conformations determines the function of the protein, with many functions depending on a protein fluctuating between multiple conformations [105–107]. This implies that understanding the evolution of protein function requires understanding how mutations alter the protein energy landscape [108,109].

The importance of an energy landscape view for understanding protein evolution can be seen in a simple engineered evolutionary trajectory. Researchers sequentially introduced mutations converting a primarily β-sheet protein into a primarily α-helical protein [110–112]. At the beginning of the trajectory, the α-helical state had a high energy and was thus unpopulated. The mutations then stabilized the α-helical conformation and destabilized the β-sheet conformation until, ultimately, the ground state of the protein switched from β-sheet to α-helical (Fig 1). A landscape view is needed to make sense of this change to the ground state, as the transition requires describing changes to both the β-sheet and α-helical conformations, as well as their relative energies at each evolutionary step. Like this engineered trajectory, natural protein evolution proceeds through mutations that tune the energy landscape [108,109,113–120].

Ancestral sequence reconstruction (ASR) is a powerful tool for revealing how energy landscapes evolve. In ASR, one uses the sequences of modern proteins and phylogenetic models to reconstruct the sequences of ancient proteins, which can then be characterized experimentally. ASR studies reveal when historical amino acid substitutions occurred and how they correlate with the acquisition of new protein features. This is an efficient means to identify residues

important for a given function [121,122]. Furthermore, ASR places protein features into a natural hierarchy: those that evolved first both set up and constrain those that evolved later [122].

ASR also provides information about how the energy landscapes of natural proteins evolve that is not readily accessible by other evolutionary analyses. For example, both ASR and co-evolutionary methods can be used to extract biophysical information from sequence alignments. ASR focuses on specific substitutions in historical proteins, allowing researchers to mechanistically dissect changes to the energy landscapes of specific family members. By contrast, co-evolutionary methods identify sites whose identities co-vary across massive sequence alignments [123,124]. This averages out sequence changes from individual evolutionary lineages, revealing architectural constraints shared by an entire protein family at the expense of detail about each family member. Likewise, experimental methods such as saturation mutagenesis and directed evolution are powerful for studies of how mutations alter energy landscapes [125,126]. But these methods deal in artificial evolutionary trajectories occurring under laboratory conditions. ASR, by contrast, provides a window into the evolution of naturally occurring proteins, which evolve in an integrated biological context and are subject to diverse evolutionary processes. Thus, ASR provides specific evolutionary information that complements other bioinformatic and experimental methods.

In this review, we discuss how ASR methods have been used to uncover the evolution of energy landscapes in naturally occurring proteins. We briefly review the methods and logic of ASR, followed by recent findings illustrating the evolution of energy landscapes and how this has shaped evolutionary trajectories. We describe how applying this lens to ASR studies helps us think about deep trends in protein evolution. Finally, we conclude with some current work that will help improve ASR as a tool for understanding the evolution of protein energy landscapes.

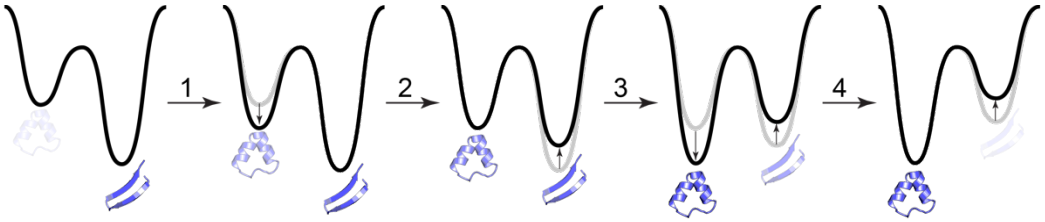

Figure 3.1. Protein evolution and energy landscapes. Cartoon of a four-mutation evolutionary trajectory that switches a small protein from a β-sheet to α-helical conformation. The black lines show the energy landscape for each genotype; the opacity of each conformation shows its population at equilibrium; the effects of mutations on the energy landscape are denoted with arrows. Mutations 1 and 3 stabilize the helical fold; mutations 2, 3, and 4 destabilize the sheet fold. Together, they switch the fold from β-sheet to α-helical.

How does ASR work?

Before diving into how ASR has been used to dissect the evolution of energy landscapes, we will provide a brief overview of the methods and statistics used in ASR. For more details, we recommend several recent reviews [127–131], as well as descriptions of software packages that explain how the calculations are done [132–134].

ASR was first proposed by Linus Pauling and Emile Zuckerkandl in 1963 [135]. They realized that analyzing the sequences of modern proteins in the context of an evolutionary tree could, in principle, allow them to reconstruct the sequences of ancient proteins. The advent of new statistical and computational approaches [136,137], massive databases of protein sequences, and cheap gene synthesis has led to an explosion in ASR studies to reconstruct the evolution of diverse protein features. These features include: enzymatic activity [116,138–144], thermodynamic stability [145–147], folding pathways [148,149], regulation [150], oligomeric state [151,152], binding specificity [153–156], fluorescent photoconversion [157,158], absorption wavelength [159–162], and many other functions [163–165].

ASR requires four steps. First, one defines a protein of interest and collects homologous sequences from diverse organisms (Fig 2A). Second, one constructs a multiple sequence

alignment (MSA) from these sequences (Fig 2B). This alignment defines which sites are homologous—that is, arose by descent—in those sequences. Third, one uses the information from the MSA to construct a phylogenetic tree describing the evolutionary relationships between the sequences (Fig 2C). Fourth, and finally, one infers the sequences of ancestors by extrapolating backwards along the inferred tree (Fig 2C).

To infer the phylogenetic tree and ancestors, most ASR studies use statistical models built on several assumptions: 1) sequences in the MSA arose by a strict bifurcation/branching process; 2) each site evolves independently; and 3) the relative probability of amino acid substitutions at each site has been the same over time. At the heart of such models is a substitution matrix encoding the probability of different evolutionary transitions (Fig 2D). These matrices are inferred from known protein sequences and thus tend to reproduce one's physiochemical intuitions. For example, in the commonly used LG matrix [166], the probability of a polar-to-polar Thr to Ser substitution is 33-times greater than a polar-to-aromatic Thr to Tyr substitution (Fig 2E). Phylogenetics software finds the phylogenetic tree that maximizes the probability of observing the MSA given the substitution model.

Ancestors are then inferred using the phylogenetic tree. For most studies, ancestors are reconstructed using the marginal probability method [137]. For each site, at each ancestral node, ASR software determines the relative probability of all ancestral scenarios given the tree and MSA. This is shown schematically in Fig 2F for an ancestral site that could plausibly be Ser or Thr. In this case, Thr is favored over Ser because the required evolutionary moves (S→T, T→T, T→T) have a higher likelihood than those required for Ser (S→S, S→T, S→T). This difference is quantified with a posterior probability: the likelihood of the most likely set of events over the

total likelihood of all events. A higher posterior probability indicates stronger support for the reconstructed amino acid at that site.

This core modeling approach is used in almost every ASR study. Additional terms may be added to better model specific evolutionary processes. These include modeling variable evolutionary rates across sites [167], using different substitution matrices for different sites in the alignment [168,169], and incorporating information from the species tree when inferring the gene tree [170]. Another key choice is whether to use a maximum likelihood approach, which infers the most plausible tree and ancestors, or a Bayesian approach, which infers a distribution of plausible trees and ancestors [171]. Given the experimental difficulties of characterizing a representative ensemble of ancestors and evolutionary trees, most ASR studies rely on a maximum likelihood approach.

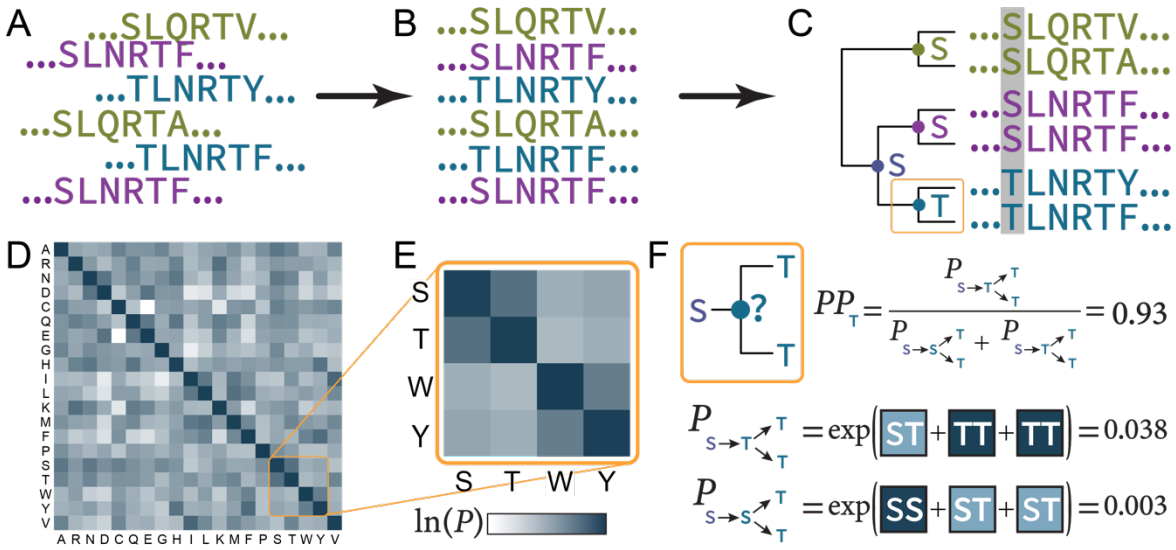

Figure 3.2: How ASR works. A) Download a diverse collection of modern sequences from online databases. B) Create a multiple sequence alignment (MSA). C) Infer a phylogenetic tree modeling the evolutionary history of the sequences in the MSA. The amino acids noted on the tree correspond to the site highlighted in gray in the MSA. D) The commonly used LG amino acid substitution matrix. Colors denote the log of the exchangeability of the amino acids denoted along x- and y-axes. E) Subset of the LG matrix showing the relative substitution probabilities of a handful of chemically similar and dissimilar amino acids. F) Calculation of the posterior probability that the amino acid at ancestor "?" was most likely Thr (PP_T). To do so, we calculate

the probability of two chains of evolutionary events: one with ancestral Thr (S→T, T→T, T→T), the other with ancestral Ser (S→S, S→T, S→T). The chain with Thr is 13 times more likely than the chain with Ser (0.038 vs. 0.003), giving a posterior probability the ancestor was Thr of 0.93. (Note: in this toy example, we assumed all branch lengths were identical).

The logic of an ASR study

The basic task in most ASR studies is to learn how a set of historical substitutions conferred a new function to an ancestral protein. A study to understand how protein feature *X* evolved would typically involve several steps [121,172]. First, find the most recent ancestor that did not have *X* (ancPreX) and the oldest ancestor that had *X* (ancPostX). *X* evolved somewhere along the evolutionary branch between these two ancestors. Then, using experiments and/or computational analyses, identify the set of mutations that conferred *X*. This is usually a subset of the total sequence differences between ancPostX and ancPreX. Finally, dissect the mechanism by which the historical mutations conferred *X*, usually by studying their effect in the historical ancPreX genetic background.

We can see how this works in practice for the evolution of protein heterocomplexes [173]. ASR has been used to trace the evolution of several multi-component protein complexes including the V0 ring of V-ATPase [174], hemoglobin [175], and other proteins [173,176]. Figure 3 shows a cartoon abstraction of these results for a hexameric complex assembled from four proteins (Fig 3, bottom right). The deepest reconstructed ancestor forms a homomeric hexamer (Fig 3, bottom left). By following how the assembly changes over subsequently more recent ancestors, one can identify the key substitutions and events that led to a highly specific modern complex. In this example, the first event was a duplication of the most ancient, homo-hexameric ancestor. Immediately after duplication, the subunits were interchangeable. This was followed by a

mutation that created a “hole” in one subunit that did not alter assembly: the subunits remained interchangeable. A second mutation created a “knob” that can only be accommodated by being adjacent to a subunit with a hole, thus conferring a specific assembly order. This process then repeated along further evolutionary branches, leading to the modern complex.

This example shows the basic logic of ASR studies, and how ASR can be used to pick apart the evolution of complicated, integrated protein features. With this strategy in mind, we can discuss how ASR has been used to learn about the evolution of protein function, specifically through the lens of energy landscapes.

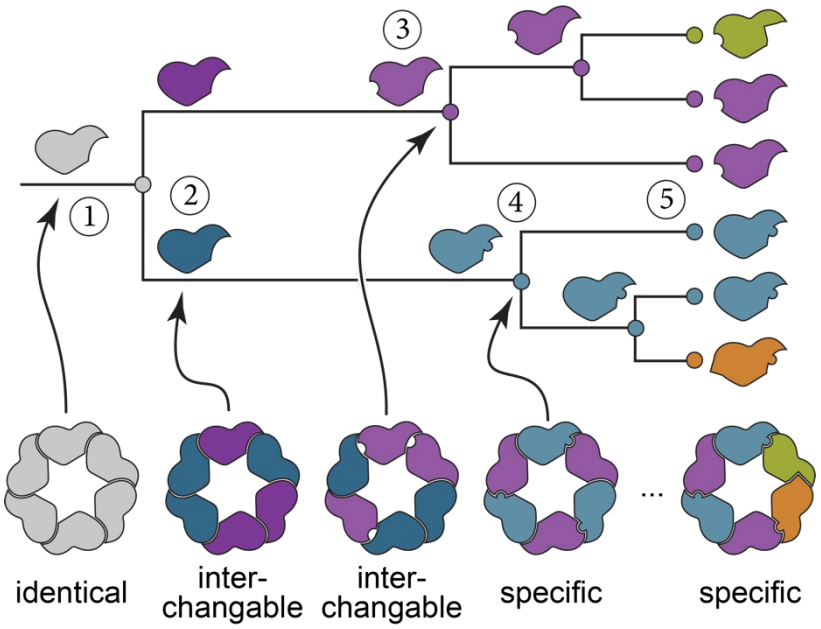

Figure 3.3: ASR can be used to trace the evolution of complicated protein features over time. This is a generalized view of the evolution of a heterohexamer, as found for multiple naturally occurring complexes [174–176]. 1) The ancestral protein forms a homohexamer. 2) The protein duplicates. The subunits have identical sequences and thus assemble in any stoichiometry and order. 3) A mutation occurs to one subunit, creating a “hole” at the interface that is compatible with any assembly order. 4) A mutation occurs on the other subunit, creating a “knob” at the interface that can only be accommodated by the “hole” on the other subunit. The proteins now form a specific hexamer with alternating subunits. 5) Further duplications and mutations allow ever-more-specific complex assembly.

Evolution of folding energy landscapes

We now examine how ASR has been used to study the evolution of energy landscapes. Folding energy landscapes are a natural place to start, as folding into the native state is a prerequisite for function in most proteins. So, what has ASR revealed about how the energy landscape controlling protein folding evolved?

One key question is how folding energy landscapes evolve in the first place. In one noteworthy example, researchers used ASR to unravel how a β-propeller fold evolved from much smaller subunits. A β-propeller is a closed repeat protein built from five tandem duplications of a short propeller-like motif [149,177]. Using lectin β-propeller evolution to model new fold emergence, Smock and colleagues resurrected an ancestral 47 aa protein encoding a single propellor-like motif [149]. This ancestral motif spontaneously forms a noncovalent pentamer in trans, mimicking the full β-propeller. Over evolutionary time, the gene encoding the monomer then duplicated, fused, and diversified to create new interfaces between subunits and, ultimately, a complete β-propeller. This evolutionary trajectory is remarkably similar to the evolutionary assembly of multi-protein complexes described in the previous section (Fig 3), suggesting similar evolutionary mechanics can operate at both the level of tertiary and quaternary structural assembly.

From the perspective of the energy landscape, one of the more intriguing aspects of this study was that they found the folding constraints changed over evolutionary time. For the ancestor, function depended on stable folding of the monomer and efficient pentamer assembly. After duplication and fusion, new constraints emerged. One of the most notable was the requirement to avoid inappropriate β-sheet formation between subunits, which leads to misfolding. The transition from a motif assembling in trans to a full β-propeller thus required

smoothing the energy landscape by destabilizing—or, at least making kinetically inaccessible—misfolded forms of the protein.

The presence of a relatively complex folding energy landscape can also open surprising opportunities for evolutionary optimization. One example comes from the folding of RNaseH [118,146,148]. This protein folds through a metastable, on-pathway intermediate. By characterizing ancestral proteins using hydrogen-deuterium exchange mass spectrometry (HDX-MS), the Marqusee group found that the formation of the major folding intermediate has been conserved over billions of years.

To understand the evolutionary implications of the preserved intermediate, the researchers traced the evolution of RNaseH proteins starting from the ancestral protein and proceeding along the branches leading to mesophilic and thermophilic descendants. They discovered that preservation of the folding intermediate has allowed RNaseH proteins to decouple the evolution of thermodynamic stability (the proportion of molecules in the folded state at equilibrium) and kinetic stability (how long it takes before a protein unfolds once it reaches the native state) (Fig 4A). While kinetic stability increased on both lineages, thermodynamic stability only increased on the thermophilic lineage (Fig 4B). This could occur because kinetic stability increased by different mechanisms along the two lineages. On the thermophilic lineage, the mutations stabilized the native state, thus increasing both thermodynamic and kinetic stability (Fig 4C-D). On the mesophile lineage, in contrast, mutations destabilized the intermediate and folding transition state energy (Fig 4C,E). This increased kinetic stability by increasing the height of the unfolding barrier without altering the proportion of molecules folded at equilibrium.

Finally, ASR has recently been used to investigate how energy landscapes can encode multiple low energy conformations. In one recent example, researchers revealed how a few mutations stabilized an alternative conformation of the protein without destabilizing the existing native conformation [114]. They found a fascinating "zigzagging" evolutionary path, beginning with an ancestral protein that had a single native state. Mutations sequentially stabilized an alternate form of the protein, eventually causing the alternate conformation to become the most stable conformation. Further mutations rebalanced the energy landscape, giving both conformations nearly identical energies. As a result, the modern protein has two interconverting native states that allow the same gene to encode multiple functions.

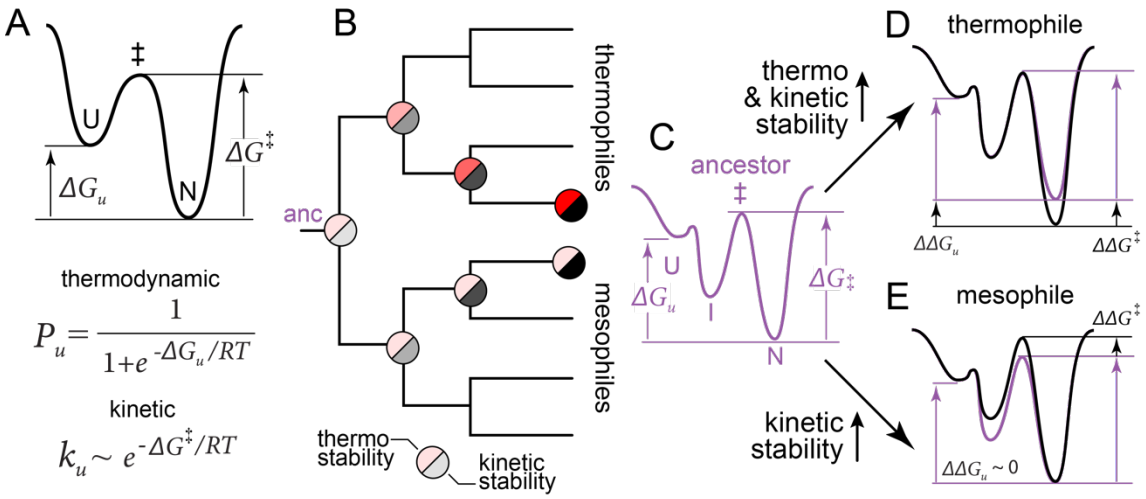

Figure 3.4. A folding intermediate decouples the evolution of thermodynamic and kinetic stability. A) Energy landscape illustrating thermodynamic stability (governed by ΔG_u) and kinetic stability (governed by $\Delta G^{\ddagger}$). B) Cartoon tree showing evolution of RNaseH. The thermophile lineage showed increased kinetic stability (gray to black) and thermodynamic stability (pink to red). The mesophile lineage increased kinetic stability without altering thermodynamic stability. C) Energy landscape of the RNaseH ancestor (under conditions favoring folding). The protein folds through an on-pathway intermediate *I*. D) Energy landscape of a modern thermophile RNaseH. Relative to the ancestor (purple), the free energy of the native state decreased, increasing both thermodynamic ($\Delta\Delta G_u$) and kinetic ($\Delta\Delta G^{\ddagger}$) stability. E) Energy landscape of a modern mesophile RNaseH. The energy of the intermediate and folding barrier increased, increasing kinetic stability without altering thermodynamic stability.

Tuning the energy landscape to confer new functions

ASR has also revealed important changes to energy landscapes within the native state ensemble. Recently, for example, ASR was used to study the evolution of new substrate specificity by the xenobiotic-degrading enzyme methyl-parathion hydrolase (MPH) [178]. MPH recently acquired the ability to act on four new organophosphate compounds. The authors resurrected the last ancestral enzyme that was unable to recognize these substrates, then identified five mutations that were necessary and sufficient to confer the new activity. These mutations had two primary effects. First, they improved the ability of the enzyme to stabilize the appropriate transition state, thus increasing the rate of catalysis (Fig 5A). Second, these mutations reduced the prevalence of non-productive modes of binding (Fig 5B). In an energy landscape view, the mutations stabilized the transition state and destabilized several non-productive binding modes. In another study, researchers found that mutations away from the active site tuned the dynamics of the protein, and thus substrate specificity [138].

The importance of mutations disrupting non-productive conformations within the landscape has proven a common feature during the evolution of new activity and function. This has even held true for ASR studies that have dissected *de novo* evolution of enzyme active sites. For several different enzymes, residues within an existing binding site were repurposed when mutations altered the conformation of the site [116,141,179]. This reorganization of the binding site pre-positioned residues to bind the reaction transition state, lowering the reaction energy barrier and conferring a low level of enzymatic activity (Fig 5A). Subsequent mutations tuned the active site to increase activity. For both the evolution of a plant flavonoid biosynthetic enzyme and a bacterial cyclohexadienyl dehydratase [141,179], the final tuning stabilized the pocket, quenching non-productive dynamics and destabilizing non-productive interactions (Fig 5B). Remarkably,

many of the required changes to cyclohexadienyl dehydratase were distant from the active site [116,180], thus demonstrating an important role for the evolution of non-active site residues in tuning the energy landscape of evolving enzymes. We also note that this work powerfully demonstrates how ASR can identify residues important for function that may not be obvious from a simple structural analysis [121,122].

Sometimes the evolution of activity requires the addition of new dynamics rather than quenched dynamics. The Matz group dissected the evolution of photoconversion activity from an ancestral GFP-like protein from corals. The ancestral GFP-like protein autocatalytically forms a green chromophore. Using ASR, the researchers identified a subset of historical substitutions that conferred the ability of the protein to photoconvert from green to red upon exposure to UV light [157]. Photoconversion requires that the chromophore rotate and pick up a nearby proton before interacting with an incoming photon. This rotation and proton pickup requires breaking an adjacent ion pair. One of the key mutations that occurred over this interval disrupted a hydrogen bond network stabilizing this ion pair, lowering the energetic barrier of rotating the chromophore and dramatically increasing the rate of photoconversion (Fig 5C) [158].

We have highlighted a few studies here, but there has been extensive work using ASR to reveal how small perturbations to the energy landscape have led to the evolution of new functions. Other examples include the evolution of a binding protein from an enzyme by massively quenching the dynamics of the protein [120]; conferring new enzyme activity by promoting oligomerization, and thus creating a new active site [151]; creating a new binding interaction by mutations promoting an induced fit mechanism [181]; mutations that tune allosteric networks [182–185]; tuning loop dynamics to alter activity [150,186], and the idea that altered heat capacity of the transition state was important for the evolution of increased enzyme activity [117].

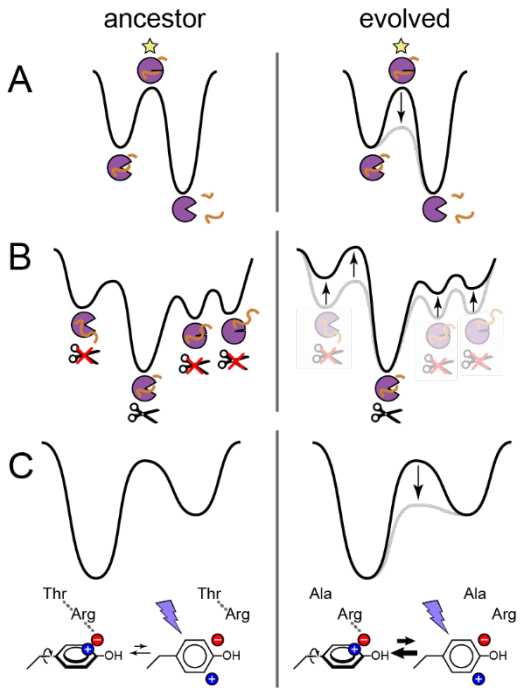

Figure 3.5. Evolution of function by altering energy landscapes. A) Icons show enzyme (purple) operating on substrate (orange). In the ancestor, the transition state energy (closed enzyme, yellow star) is too high to allow catalysis; mutations that lower the energy of the transition state convert the ancestral protein into an enzyme. B) An enzyme binds to a substrate in multiple conformations. One is productive (denoted with scissors); the other are non-productive (red "×"). Mutations that destabilize the non-productive conformations increase the bulk rate of catalysis. C) The slow step in the photoconversion of a Kaede GFP-like chromophore is breaking an ion pair near the chromophore (blue "+" and red "-"). A historical Thr to Ala mutation at a site away from the chromophore disrupted a polar network stabilizing the ion pair, thus increasing the rate of photoconversion.

Energy landscapes constrain evolution

ASR has also revealed ways in which protein energy landscapes shape evolution. One example comes from our own work. We used ASR to study the evolution of a new pro-inflammatory function by the mammalian protein S100A9 [187,188]. Reverting a functionally important site in the human protein to its ancestral phenylalanine disrupted the function of the protein. The reversion creates a new Phe/Phe contact that stabilizes a non-functional form of the protein. In wildtype S100A9, the non-functional form of the protein is quite close in energy to the native state—so close that a single new Phe/Phe interaction makes it the new native state (Fig

6A). A co-evolutionary analysis of S100A9 proteins from across mammals revealed that Phe can be found at either site in the Phe/Phe pair, but almost never at both sites together. This suggests the protein has been evolving to avoid this deleterious contact, and that the energy landscape constrains how the protein can evolve.

Other studies have revealed similar constraints [189–191]. For example, the glucocorticoid receptor in bony vertebrates evolved ligand specificity through a conformational shift in a helix bordering the binding site. This was enabled by "permissive" mutations that stabilized the helix in its new conformation. Although many mutations can stabilize the protein, only a few can act as permissive mutations. This is because many of the stabilizing mutations shift the energy landscape to favor the active form even in the absence of ligand, thus making the receptor constitutively active [189].

Energy landscapes open and close evolutionary trajectories

The Phe/Phe interaction from above is an example of the broad phenomenon of *epistasis*, where the effect of a mutation changes depending on the presence of other mutations (Fig 6B). Epistasis profoundly alters evolutionary outcomes by opening and closing evolutionary trajectories. Fig 6B shows a mutant cycle with two mutations a→A and b→B. The a→A mutation disrupts function on its own, but can be tolerated after the b→B mutation. This means the evolutionary trajectory ab→aB→AB is favored over the ab→Ab→AB trajectory. Epistasis introduces contingency into evolution [189,192,193]: a→A can only occur after the permissive b→B mutation. Epistasis also causes entrenchment [159,193,194]: once the restrictive a→A mutation occurs in the aB background, it prevents the reversion B→b.

Epistasis can arise directly from protein energy landscapes. One way, well documented in a variety of ASR studies, is via the stability of the native state [190,191,195,196] (Fig 6C). A mutation that compromises protein stability may not be tolerated unless it is preceded by a different mutation that stabilizes the protein (Fig 6C). Such stability tradeoffs often occur during the evolution of new functions. Function-switching mutations often compromise protein stability by creating unsatisfied interactions that confer binding specificity or, in the case of enzymes, stabilize a transition state; these destabilizing effects are often offset by stabilizing mutations [190,191,195,196].

ASR has also revealed more subtle ways that mutations can perturb energy landscapes, thus changing the accessibility of evolutionary trajectories. Figure 6D-F shows a cartoon landscape that starts with a weak interaction between two molecules (denoted with colored circles and a purple bar). This weak interaction can occur via two, initially isoenergetic, conformations (γ and δ). If mutations "1" and "2" occur in the purple molecule, they increase the overall favorability of the interaction by sequentially stabilizing δ and destabilizing γ (Fig 6D). Mutations "3" and "4" promote the same interaction, but instead stabilize γ over δ (Fig 6E). Starr and colleagues observed this phenomenon for the evolution of transcription factors interacting with specific DNA response elements [197]. Using a combination of ASR and deep mutational scanning, they found that there were several structurally different ways for an ancestral transcription factor to evolve to bind specific DNA sequences. Commitment to one binding model excluded the other binding model. This leads to profound epistasis. In the evolutionary trajectory shown in Fig 6E, mutation "3" stabilizes the interaction; however, the same mutation destabilizes the interaction in Fig 6F because mutation "1" has already committed the interaction

to favor conformation δ rather than γ. The initial mutation selecting one conformation or the other entrenches that particular outcome.

These examples demonstrate "pairwise" epistasis between two mutations; however, ASR studies have also revealed examples of "high-order" epistasis between three or more mutations [157,178,191,194]. In three-way epistasis, for example, the effect of three mutations placed together cannot be predicted from their individual effects combined with any pairwise epistasis [198–200]. Such high-order interactions arise naturally on energy landscapes that populate more than two conformations [201]. In the ASR study described above, where the authors dissected the evolution of new enzymatic activity in MPH [178], the authors generated all 2^5 combinations of the five mutations they identified as important of the new activity. They found extensive high-order epistasis between the mutations, mediated by subtle changes in structure and binding energy. This highly constrained the evolution of MPH activity. Out of the 120 (5!) possible paths through this functional landscape, only 19 were accessible between the ancestral and extant MPH, underscoring the importance of mutation order in evolution. This study joins a wealth of other ASR studies showing how subtle biophysical changes can profoundly alter evolutionary outcomes by inducing epistasis [189–191,197,202].

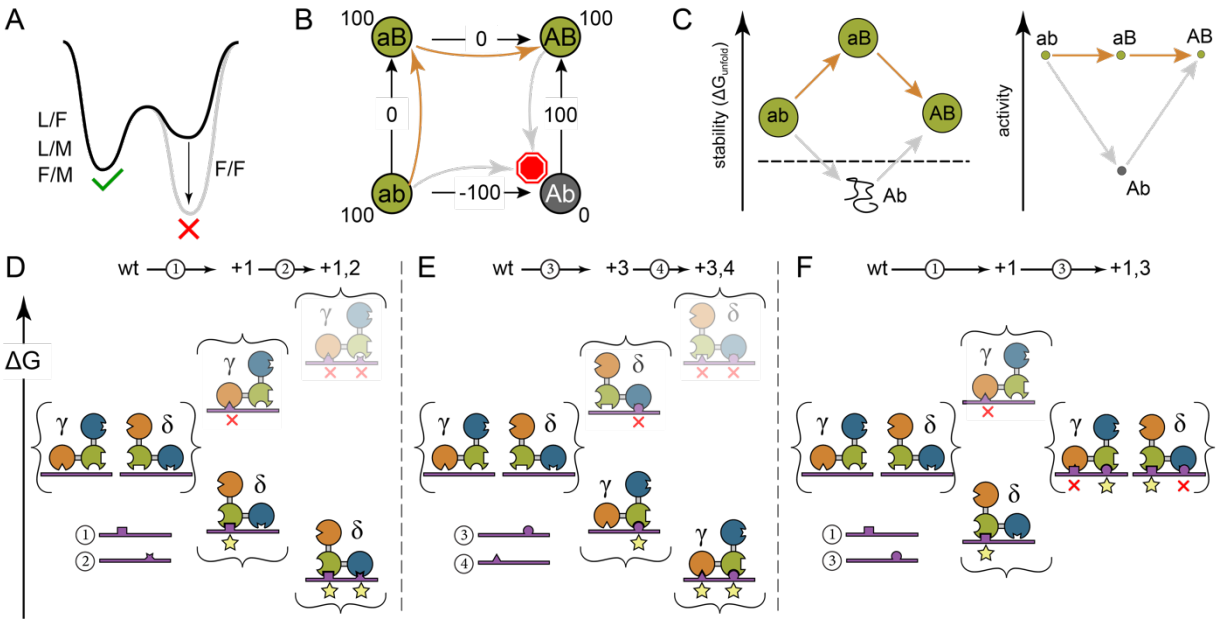

Figure 3.6. Epistasis arising from energy landscapes. A) The S100A9 energy landscape constrains evolution; Phe is tolerated at either site, but placing Phe at both sites stabilizes a non-functional form of the protein. B) Epistasis between mutations a→A and b→B shapes evolutionary trajectories. Genotypes ab, aB, and AB have the same activity (100; green); genotype Ab is nonfunctional (0; gray). Only the ab→aB→AB evolutionary trajectory is accessible without compromising function (orange arrows). Mutation a→A cannot be tolerated without the b→B mutation (bottom gray arrow, stop sign). Reversion of B→b cannot be tolerated after the a→A mutation (right gray arrow; stop sign). C) One biophysical explanation for the epistasis in panel B. Mutations a→A and b→B have opposite, additive effects on protein stability, leading to epistasis in activity. Genotype Ab is unfolded (left); therefore, there is less active protein in the cell and activity is low (right). D) Evolution of a new interaction between two macromolecules leads to epistasis. The purple molecule acquires mutations promoting interaction with the molecule drawn as colored circles. Arrows and letters above the panel indicate the evolutionary trajectory. The wildtype purple genotype allows two isoenergetic conformations (γ and δ). Mutations "1" and "2" to the purple molecule stabilize conformation δ (yellow stars) and destabilize conformation γ (red ×), this increases the affinity of the interaction. E) Identical to panel D, except mutations "3" and "4" to the purple molecule stabilize γ and destabilize δ. F) Epistasis mediated by the energy landscape. When introduced after "1", mutation "3" destabilizes the interaction because it "1" and "2" have opposite effects on γ and δ.

Increasingly thermostable ancestral states: an artifact of the energy landscape?

Viewing protein evolution from the perspective of an energy landscape may also shed light on some puzzling observations from ASR studies. One such observation is that resurrected ancestral proteins often exhibit increasing thermostability as one moves deeper into the past (Fig 7A) [119,142,145,147,203–210]. This makes reconstructed proteins robust and opens the door for aggressive engineering approaches that would otherwise result in primarily "dead" proteins [131,211–216].

The origins of this trend, however, remain unclear. Some have argued that the trend to higher stability in the past arises because ancestral environments were hotter, necessitating more stable ancient proteins [217,218]. This would make ASR a useful tool to study ancient environments [145,219]. One difficulty with this view is that temperature conditions varied widely over time and geographical location, making the uniformity of these trends across all lineages difficult to rationalize [220].

Others have argued that the trend towards hyperstability is an artifact of the methods used to reconstruct ancestors [221–224]. One possible cause of artifactually high stability of ancestral proteins would be an inappropriate introduction of consensus residues. Consensus proteins—where one substitutes the most common amino acid seen in a multiple sequence alignment at each site in the protein—are consistently hyperstable [221]. If the same stabilizing amino acid was acquired convergently on multiple lineages, ASR methods could incorrectly infer that the ancestor had that amino acid [224]. As a result, ancestral proteins would have, on average, an excess number of stabilizing amino acids and would thus prove artificially stable. This effect could also arise from phylogenetic model violation. Most common methods assume that the frequencies of amino acids at each site have not changed over time; if this is not true, the reconstructed ancestors could be biased towards amino acids with higher frequency in the alignment [222].

The evidence that ASR converges on consensus sequences remains mixed. Not all reconstructed ancestors appear consensus-like [221,223]. Further, it is not clear that adding a handful of consensus mutations would be sufficient to confer hyperstability. Sternke and colleagues compared the measured effects of mutations across a database of proteins to their amino acid frequencies in the proteomes. There was low correlation between the relative stabilizing effect of the mutations and the frequency of amino acids, suggesting the consensus explanation may not be sufficient to explain the stabilization effect [221]. Further, given the degree of epistasis in proteins, it is not clear that a consensus mutation, introduced by itself, would be universally stabilizing across all backgrounds. This would be especially true for evolutionary changes in function that induce evolutionary "Stokes shifts"—a reorganization of residue preference across sites after a large-effect substitution [225,226]. The origins of ancestral hyperstability thus remain unclear.

With the protein energy landscape in view, we propose another way ASR could artificially increase protein stability: evolution towards a *consensus landscape* rather than a *consensus sequence*. This idea is tentative and must be tested. The idea has three parts: 1) As most proteins evolve, their native state is under purifying selection and thus remains relatively unchanged. The residues encoding interactions in the native state will evolve relatively slowly. 2) In addition to the native conformation, protein energy landscapes possess relatively low-energy non-native conformations (Fig 7B). These conformations have no functional constraints—other than not becoming so populated they disrupt function—and thus are randomly formed and destroyed as the protein sequence evolves [227]. 3) When an ancestor is reconstructed, it will have higher quality signal for residues encoding the native conformation than the non-native conformations. Because the sequences and identities of the non-native interactions change

relatively rapidly over time, they are “smeared out” by the reconstruction, leading toward an “average” landscape dominated by the native structure (Fig 7C,D). The net result is a smoother energy landscape and a more dominant contribution to stability by the native state. This effect would become greater the further back one went in time, as one is averaging over a larger number of non-native conformations.

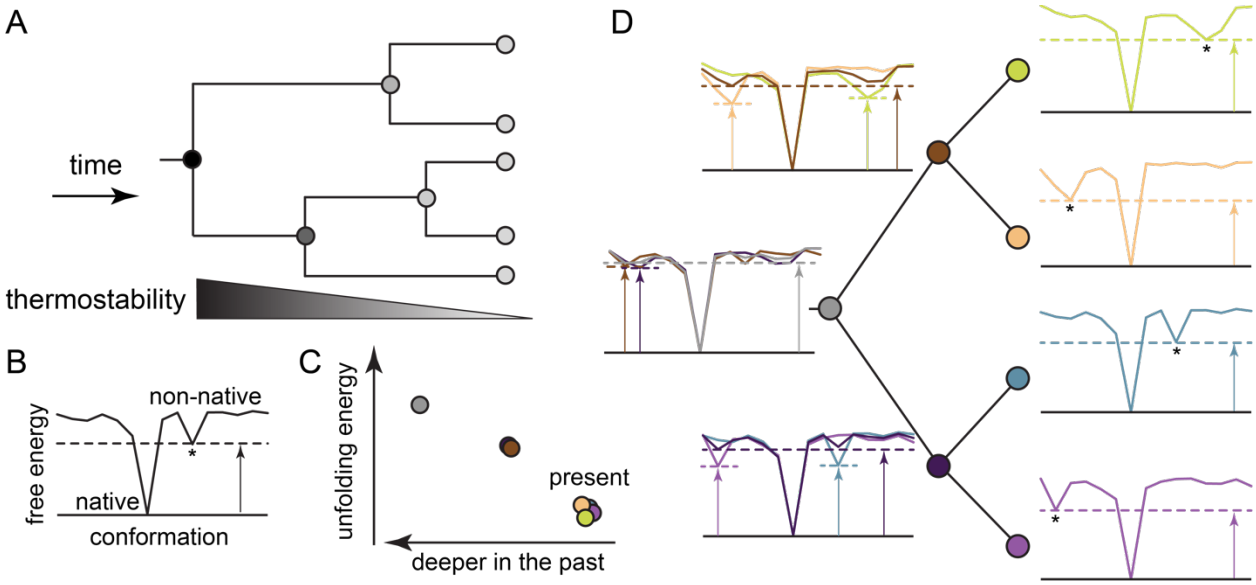

Figure 3.7. Evolution of consensus landscapes as a mechanism for ancestral stabilization? A) ASR studies have found that more ancient ancestors, on most lineages, have higher thermostability than later ancestors. Thermostability is represented as a gradient from high (dark) to low (light). B) An energy landscape with a protein that has a single high-energy non-native conformation. The arrow indicates the energy difference between the native state and the lowest energy excited state (star). C) A hypothetical trend in protein stability, with ancestral colors corresponding to the tree in C. D) Modern proteins on the right have the same native state, but different excited states. Because of this, the evolutionary signal is stronger for the native state than the excited states. When ancestors are reconstructed, the physical interactions encoding the native state are correctly inferred, leading to an accurate native state conformational energy. The evolutionary signal for the excited states is weak; therefore, the interactions stabilizing specific excited states are not accurately inferred. Ancestors have an “average” set of non-native interactions that is incompatible with specific non-native conformations. Thus, reconstructed ancestors maintain the native conformation energy, but sequentially average out the energies of non-native conformations, leading to a larger energy difference between the native and non-native conformations the deeper one goes back in time.

Limitations and future directions

While ASR has proven powerful, there are limitations to the approach. Some of these limitations are intrinsic. Evolutionary models will always be approximations of a complicated historical process. Further, some evolutionary events would require sequences from species that went extinct and thus have no modern sequences to include in an MSA.

Some limitations to ASR methods can, however, be addressed and improved. Substitution models (Fig 2D) are one important area for improvement. The most popular models assume that all sites in the protein have the same substitution probabilities and, on long time scales, converge to the same amino acid preferences [222]. It is not obvious, however, that one should treat the evolution of a site in the core of a protein with the same model one treats a surface residue, nor that the preferences of amino acids at a site might change over time. Models have been developed that use different substitution models for different classes of sites. One model, for example, uses six matrices to treat sites classified based on their solvent accessibility and secondary structures [168]. Other work has been done to empirically define the amino acid preferences at sites using deep mutational scans [228]. Others have proposed using non-stationary models of amino acid frequencies, thus allowing amino acid preference to change over time [222,229]. Despite having clear utility for ASR studies [230]—particularly for biophysicists dissecting the detailed features of energy landscapes—many of these models have seen limited use in practice [230,231]. This is likely because such models have not been incorporated into mainstream phylogenetics software and require coding skill and/or detailed phylogenetics knowledge to use.

Other aspects of the models that can be improved are how the models capture variations in evolutionary rate across sites and time, explicit models of insertion and deletion, better methods to bring in outside information such as a species tree at the time of inference, attempts

to treat co-variation between sites, and better tree-search algorithms. These are active areas of development within the phylogenetics community [232,233]: continued improvements will almost certainly lead to better reconstructed ancestors for biophysical study. That said, it is critical for researchers to keep in mind that ASR studies can only confidently be done on protein families that adhere reasonably well to the assumptions of current phylogenetic models.

Conclusion

ASR has proven a powerful means to access how protein energy landscapes evolve. By separating the evolution of protein features in time, ASR provides a nuanced view of how mutations alter energy landscapes during the evolution of protein function. Further, ASR has revealed how constraints due to the energy landscape have shaped—and continue to shape—the evolution of protein function. Continued methodological development promises to make the tool even more useful for biophysicists. As the approach is applied to ever more protein families, it will continue to reveal both how proteins acquired their amazingly diverse functions, as well as general principles to understand and, maybe someday, predict protein evolution.

Bridge to chapter IV

Chapter III covers an extensive body of literature consisting of both the biophysics of proteins and also the power of Ancestral Sequence Reconstruction as an evolutionary tool. We also provide perspectives on future work in the field and how we could potentially intertwine the protein energy landscape with ASR studies. Chapter III wraps up the technique overview and exploration for this dissertation, while Chapter IV will close out the rest of the document with overall conclusions from each chapter and postulate on some future directions for the field of protein evolution as a whole.

CHAPTER IV

CONCLUSION AND FUTURE DIRECTIONS

Overall, this dissertation covers a brief look into the capabilities of biophysical perspectives in evolutionary biochemistry studies. Linking powerful experimental and computational tools can open a unique glimpse into the features underlying important patterns in protein evolution. In this work, we introduced the potential implications of biophysical lenses in evolution through covering some key innovations and techniques that have been used in the past few decades in this field. Machine learning, for example, provides an avenue for studying protein evolution with much promise, but needs to be further refined before robust results can be obtained considering epistasis. However, there are promising experimental frameworks that can be leveraged to study protein evolution and the biophysical features that guide that process. This work focused specifically on the experimental study of protein sequence space and

understanding how proteins navigate this space through changes in biophysical characteristics. Additionally, this work covered a review of Ancestral Sequence Reconstruction and how a new perspective through biophysics could enhance the utility of the technique in evolutionary studies.

In chapter II, we experimentally explore sequence space and the underlying biophysics that determine the functional robustness of that space using a naturally evolved fluorescent protein. This work showed that local volumes of protein sequence space are subject to change in their functional composition over evolutionary time, shifting from mutationally robust local spaces to extremely susceptible local spaces. We also explored a few potential biophysical features that may be guiding these changes in local neighborhood robustness. Surprisingly, global stability of the protein did not directly correlate with mutational robustness. Altogether, this work in chapter II demonstrated the utility of using high-throughput experimental characterization across evolutionary steps in protein sequence space, and provides a framework for future work in other protein contexts.

In chapter III, we detail the design and implementation of Ancestral Sequence Reconstruction, a powerful tool for recapitulating ancient proteins and their evolutionary intermediates. We also offer insight into the usage of a new biophysical lens that could expand our interpretations from ASR studies via the protein energy landscape. Moving forward in the future, we propose that this protein energy landscape lens coupled with ASR could provide a unique look into the biophysical underpinnings of evolution across a likely historical trajectory.

While there is still much development to be done in the field of protein evolution, proper consideration of the biophysical principles that underlie the evolutionary process is a key component that cannot be overlooked. Experimental and computational tools integrating this aspect into evolutionary models and frameworks should be a focus in the field moving forward,

specifically with relevance to protein sequence space. While protein sequence space is too vast to exhaustively explore, a focused and thorough look into relevant areas of sequence space over evolutionary time can shed light on answers to many of the questions posed in this dissertation. Uncovering the functional connectivity of protein sequence space and the biochemical and biophysical traits that determine that connectivity will be foundational to our understanding of protein evolution in the future.

APPENDIX

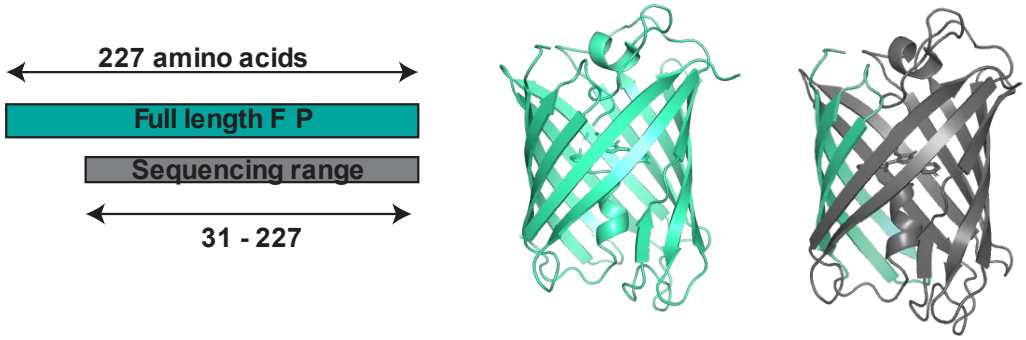

Fig APP1: The first 30 N-terminal amino acids are not included in the library mutagenesis Cartoon depiction of the amino acids included in the library mutagenesis for each fluorescent protein compared to the full length of the protein (left). Structures showing the mutagenized region of each protein library (right). The dark gray color represents the region of amino acids that are included in the library mutagenesis.

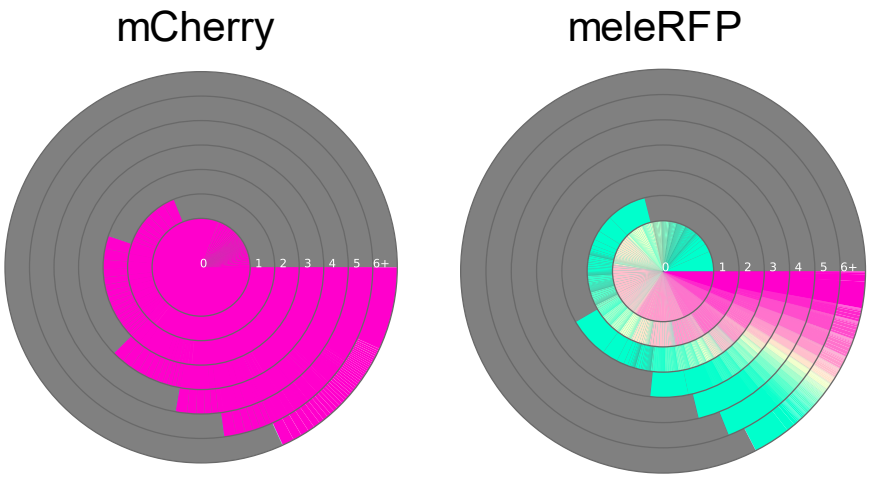

Fig APP2: Unrelated fluorescent protein neighborhoods are not mutationally robust Local neighborhoods of mCherry and meleRFP. Consistent with the main text neighborhood diagrams, the gray fractions represent non-fluorescent mutant proteins. Colored fractions indicate measured colors of mutant proteins. The center ring includes proteins with no mutations, while each subsequent ring increases in mutations up to 6+.

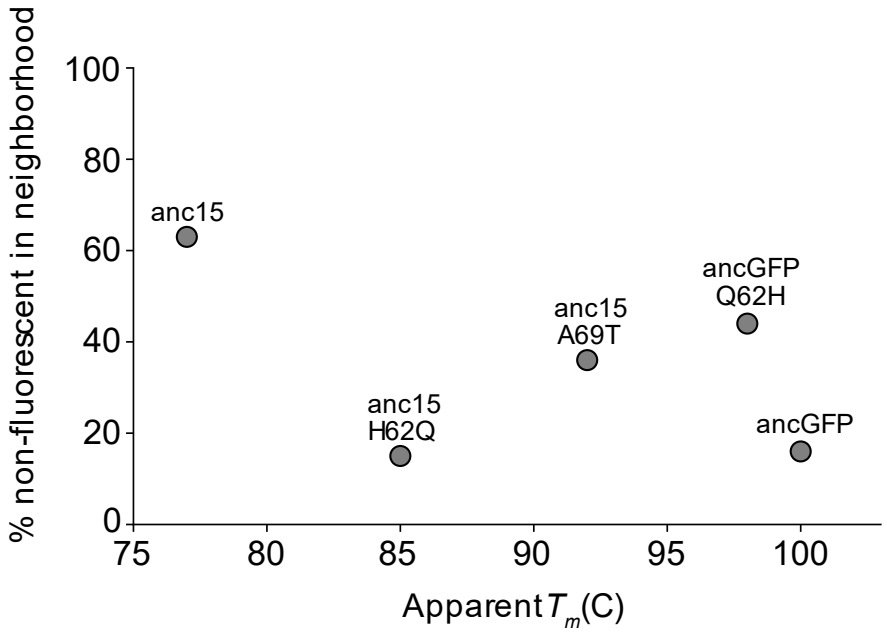

Fig APP3: Thermal stability measurements correspond to chemical denaturation results
Thermal stability of fluorescent proteins as measured by fluorescence signal on a Jasco (XX) CD spectrometer. The apparent Tm was determined by the midpoint of the folded to unfolded transition for each protein (don't know a good way of saying this). Y-axis corresponds to the maximum % non-fluorescent proteins observed in a given protein's neighborhood.

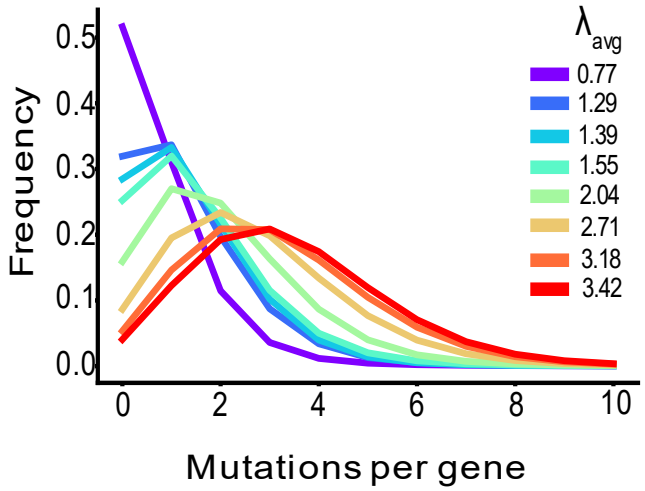

Fig APP4. Library mutagenesis yields protein libraries with different mutation rates
Calibrated mutation rates for all 8 mutant libraries generated from error-prone PCR. Each color represents a different library, with the corresponding average mutation rate on the right side of the graph. Mutations in each library closely follow a Poisson distribution.

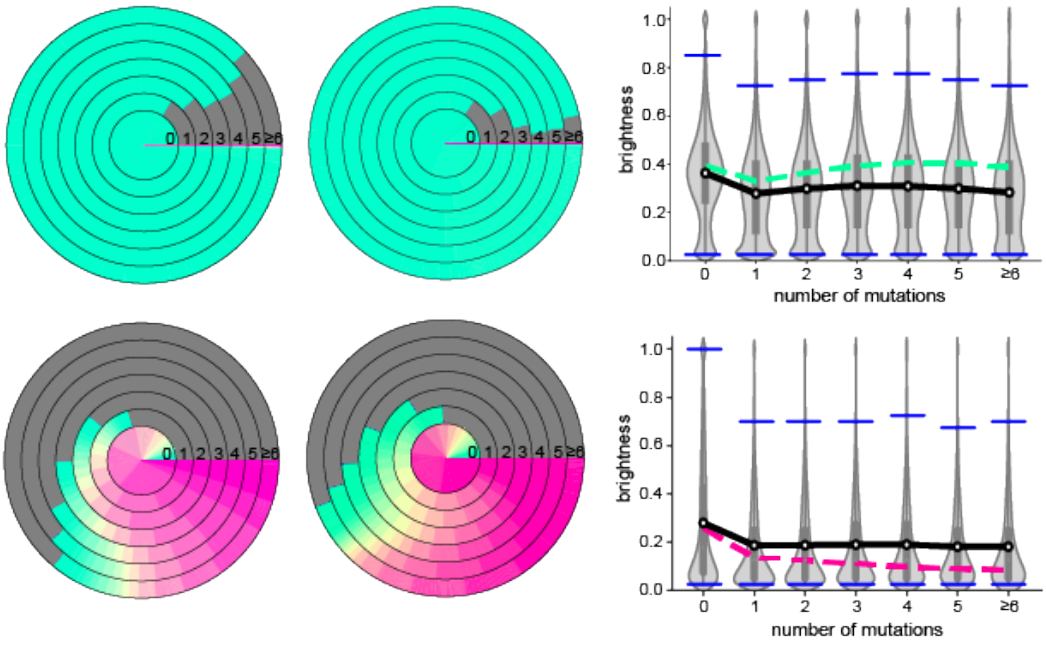

Fig APP5. Neighborhood characterization is similar across day-to-day variation
Neighborhood comparisons of ancGFP (top) and anc15 (bottom) from two separate days of preparing the same mutant library stocks. anc15 reps were both photoconverted for 15 minutes prior to characterization on the cytometer. The green and pink dashed lines in the brightness graphs denote the average brightness of proteins with a given mutation from the first rep of ancGFP or anc15. The black lines are the average brightness across the neighborhood for the second rep of either ancGFP or anc15.

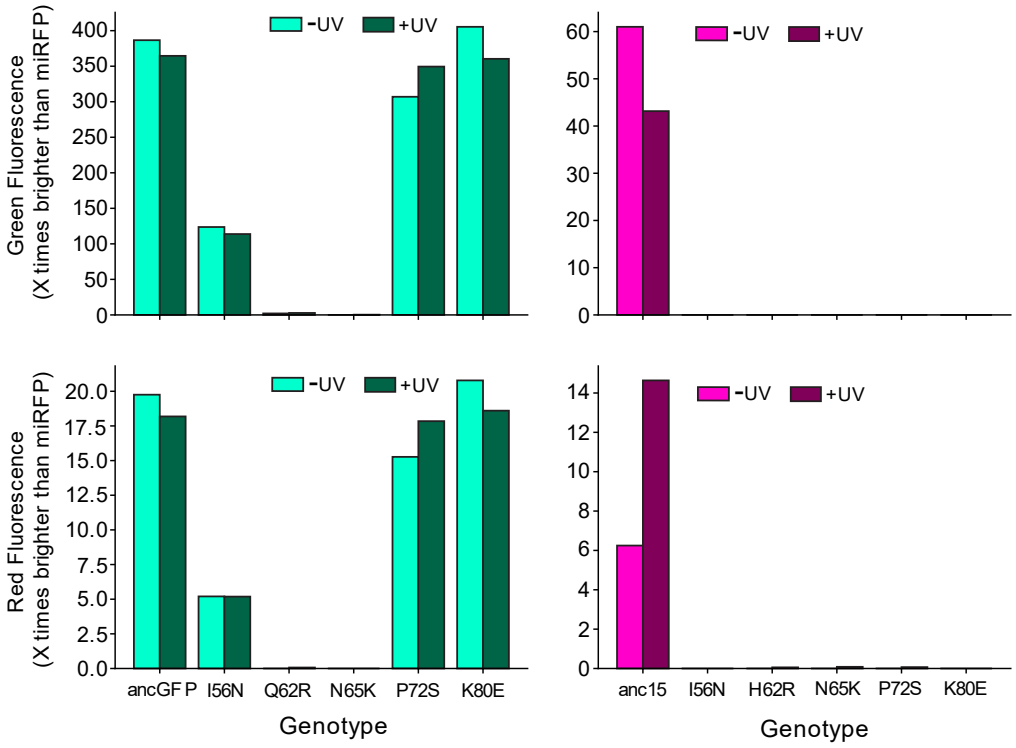

Fig APP6. ancGFP mutants are more robust than anc15 mutants in red and green wavelengths. Fluorescence emission of ancGFP, anc15, and their associated point mutants used in validating the ancGFP and anc15 neighborhood data. Emission is measured at 510 nm (top row) and 585 nm (bottom row). Emission is shown relative to the control fluorescence of the attached and unmutated fusion miRFP703 construct. Bars are shown for emission either with no UV exposure or 30 minutes of continuous UV exposure.

Genotype	C_m (Molar)	ΔG_{H2O}	m-value
ancGFP	3.25	1.65	-0.51
ancGFP Q62H	2.90	1.40	-0.48
ancGFP T69A	3.11	1.75	-0.56
anc15	2.03	3.06	-1.51
anc15 H62Q	2.24	2.31	-1.03
anc15A69T	2.40	1.82	-0.76

Table APP1. Stability fit parameters for all fluorescent proteins. Fit parameters for the 2-state unfolding model fit in figure 7A. The Cm values for each protein shown in the second column are the same values used for figure 7B. Y-intercept (dG) and slope (m-value) are also shown in the last two columns.

REFERENCES CITED

1. McEntee, K., Weinstock, G. M. & Lehman, I. R. Binding of the recA protein of Escherichia coli to single- and double-stranded DNA. *Journal of Biological Chemistry* **256**, 8835–8844 (1981).

2. Stock, A., Chen, T., Welsh, D. & Stock, J. CheA protein, a central regulator of bacterial chemotaxis, belongs to a family of proteins that control gene expression in response to changing environmental conditions. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1403–1407 (1988).

3. Hartwell, J. *et al.* Phosphoenolpyruvate carboxylase kinase is a novel protein kinase regulated at the level of expression. *Plant J* **20**, 333–342 (1999).

4. Cianciotto, N. P., Cornelis, P. & Baysse, C. Impact of the bacterial type I cytochrome *c* maturation system on different biological processes. *Molecular Microbiology* **56**, 1408–1415 (2005).

5. Hines, J. K., Fromm, H. J. & Honzatko, R. B. Structures of Activated Fructose-1,6-bisphosphatase from Escherichia coli. *Journal of Biological Chemistry* **282**, 11696–11704 (2007).

6. Horak, R. E. A. *et al.* Ammonia oxidation kinetics and temperature sensitivity of a natural marine community dominated by Archaea. *ISME J* **7**, 2023–2033 (2013).

7. Caffarri, S., Tibiletti, T., Jennings, R. & Santabarbara, S. A Comparison Between Plant Photosystem I and Photosystem II Architecture and Functioning. *CPPS* **15**, 296–331 (2014).

8. Buehner, M., Ford, G. C., Moras, D. & Olsen, K. W. D-Glyceraldehyde-3-Phosphate Dehydrogenase: Three-Dimensional Structure and Evolutionary Significance.

9. Heuser, J. E. & Kirschner, M. W. Filament organization revealed in platinum replicas of freeze-dried cytoskeletons. *The Journal of cell biology* **86**, 212–234 (1980).

10. Binarová, P. & Tuszynski, J. Tubulin: Structure, Functions and Roles in Disease. *Cells* **8**, 1294 (2019).

11. Davies, K. M. *et al.* Macromolecular organization of ATP synthase and complex I in whole mitochondria. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 14121–14126 (2011).

12. Biggar, W. D. & Sturgess, J. M. Role of lysozyme in the microbicidal activity of rat alveolar macrophages. *Infect Immun* **16**, 974–982 (1977).

13. Patel, P. H., Kawate, H., Adman, E., Ashbach, M. & Loeb, L. A. A Single Highly Mutable Catalytic Site Amino Acid Is Critical for DNA Polymerase Fidelity. *Journal of Biological Chemistry* **276**, 5044–5051 (2001).

14. Jayaraman, B., Fernandes, J. D., Yang, S., Smith, C. & Frankel, A. D. Highly Mutable Linker Regions Regulate HIV-1 Rev Function and Stability. *Sci Rep* **9**, 5139 (2019).

15. Bianchi, M., Borsetti, A., Ciccozzi, M. & Pascarella, S. SARS-Cov-2 ORF3a: Mutability and function. *International Journal of Biological Macromolecules* **170**, 820–826 (2021).

16. Shinkai, A., Patel, P. H. & Loeb, L. A. The Conserved Active Site Motif A of Escherichia coliDNA Polymerase I Is Highly Mutable. *Journal of Biological Chemistry* **276**, 18836–18842 (2001).

17. van der Meer, J.-Y., Biewenga, L. & Poelarends, G. J. The Generation and Exploitation of Protein Mutability Landscapes for Enzyme Engineering. *ChemBioChem* **17**, 1792–1799 (2016).

18. Tan, B.-C. *et al.* Identification of an Active New *Mutator* Transposable Element in Maize. *G3 Genes|Genomes|Genetics* **1**, 293–302 (2011).

19. Longya, A. *et al.* Gene Duplication and Mutation in the Emergence of a Novel Aggressive Allele of the *AVR-Pik* Effector in the Rice Blast Fungus. *MPMI* **32**, 740–749 (2019).

20. Cline, J. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Research* **24**, 3546–3551 (1996).

21. Hartl, D. Compensatory Nearly Neutral Mutations: Selection without Adaptation. *Journal of Theoretical Biology* **182**, 303–309 (1996).

22. Malinin, N. L. *et al.* A point mutation in KINDLIN3 ablates activation of three integrin subfamilies in humans. *Nat Med* **15**, 313–318 (2009).

23. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *Journal of Molecular Biology* **369**, 1318–1332 (2007).

24. Valbuena, F. M. *et al.* A photostable monomeric superfolder green fluorescent protein. *Traffic* **21**, 534–544 (2020).

25. Kunst, C. B., Mezey, E. & Patterson, D. Mutations in S0D1 associated with amyotrophic lateral sclerosis cause novel protein interactions. **15**, (1997).

26. Blum, M., Waldner, M. & Gisi, U. A single point mutation in the novel PvCesA3 gene confers resistance to the carboxylic acid amide fungicide mandipropamid in Plasmopara viticola. *Fungal Genetics and Biology* **47**, 499–510 (2010).

27. Lyons, D. & Lauring, A. Mutation and Epistasis in Influenza Virus Evolution. *Viruses* **10**, 407 (2018).

28. Horne, J. & Shukla, D. Recent Advances in Machine Learning Variant Effect Prediction Tools for Protein Engineering. *Ind. Eng. Chem. Res.* **61**, 6235–6245 (2022).

29. Austin, H. P. *et al.* Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U.S.A.* **115**, (2018).

30. Łuksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).

31. McLean, H. Q. *et al.* Interim Estimates of 2022–23 Seasonal Influenza Vaccine Effectiveness — Wisconsin, October 2022–February 2023. *MMWR Morb. Mortal. Wkly. Rep.* **72**, 201–205 (2023).

32. Price, A. M. Influenza Vaccine Effectiveness Against Influenza A(H3N2)-Related Illness in the United States During the 2021–2022 Influenza Season.

33. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465**, 922–926 (2010).

34. Schreiber, G., Buckle, A. M. & Fersht, A. R. Stability and function: two constraints in the evolution of barstar and other proteins. *Structure* **2**, 945–951 (1994).

35. Studer, R. A., Christin, P.-A., Williams, M. A. & Orengo, C. A. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2223–2228 (2014).

36. Thomas, V. L., McReynolds, A. C. & Shoichet, B. K. Structural Bases for Stability–Function Tradeoffs in Antibiotic Resistance. *Journal of Molecular Biology* **396**, 47–59 (2010).

37. Barrozo, A., Duarte, F., Bauer, P., Carvalho, A. T. P. & Kamerlin, S. C. L. Cooperative Electrostatic Interactions Drive Functional Evolution in the Alkaline Phosphatase Superfamily. *J. Am. Chem. Soc.* **137**, 9061–9076 (2015).

38. Franzosa, E. A. & Xia, Y. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution* **26**, 2387–2395 (2009).

39. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13567–13572 (2015).

40. Zimmerman, M. I. *et al.* Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Cent. Sci.* **3**, 1311–1321 (2017).

41. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801–807 (2014).

42. Koch, P. *et al.* Optimization of the antimicrobial peptide Bac7 by deep mutational scanning. *BMC Biol* **20**, 114 (2022).

43. Dadonaite, B. *et al.* A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell* **186**, 1263-1278.e20 (2023).

44. Bloom, J. D. An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit. *Molecular Biology and Evolution* **31**, 1956–1978 (2014).

45. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families: Generative Models for Protein Fold Families. *Proteins* **79**, 1061–1078 (2011).

46. Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nat Commun* **12**, 2403 (2021).

47. Hawkins-Hooker, A. *et al.* Generating functional protein variants with variational autoencoders. *PLoS Comput Biol* **17**, e1008736 (2021).

48. Greener, J. G., Moffat, L. & Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci Rep* **8**, 16189 (2018).

49. Rohlhill, J., Sandoval, N. R. & Papoutsakis, E. T. Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated *Escherichia coli* Growth on Methanol. *ACS Synth. Biol.* **6**, 1584–1595 (2017).

50. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Current Opinion in Structural Biology* **69**, 131–141 (2021).

51. Field, S. F. & Matz, M. V. Retracing Evolution of Red Fluorescence in GFP-Like Proteins from Faviina Corals. *Molecular Biology and Evolution* **27**, 225–233 (2010).

52. Ugalde, J. A., Chang, B. S. W. & Matz, M. V. Evolution of Coral Pigments Recreated. *Science* **305**, 1433–1433 (2004).

53. Kim, H. *et al.* A Hinge Migration Mechanism Unlocks the Evolution of Green-to-Red Photoconversion in GFP-like Proteins. *Structure* **23**, 34–43 (2015).

54. Fraser, J. S. *et al.* Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 16247–16252 (2011).

55. Dobo, A. & Kaltashov, I. A. Detection of Multiple Protein Conformational Ensembles in Solution via Deconvolution of Charge-State Distributions in ESI MS. *Anal. Chem.* **73**, 4763–4773 (2001).

56. Shehu, A., Kavraki, L. E. & Clementi, C. Multiscale characterization of protein conformational ensembles. *Proteins* **76**, 837–851 (2009).

57. Wen, J., Chen, X. & Bowie, J. U. Exploring the allowed sequence space of a membrane protein. *Nat Struct Mol Biol* **3**, 141–148 (1996).

58. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**, 678–687 (2005).

59. Shakhnovich, B. E., Deeds, E., Delisi, C. & Shakhnovich, E. Protein structure and evolutionary history determine sequence space topology. *Genome Res.* **15**, 385–392 (2005).

60. Taylor, S. V., Walter, K. U., Kast, P. & Hilvert, D. Searching sequence space for protein catalysts. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10596–10601 (2001).

61. Nussinov, R., Tsai, C.-J. & Jang, H. Protein ensembles link genotype to phenotype. *PLoS Comput Biol* **15**, e1006648 (2019).

62. Zheng, J., Guo, N. & Wagner, A. Selection enhances protein evolvability by increasing mutational robustness and foldability. *Science* **370**, eabb5962 (2020).

63. Field, S. F., Bulina, M. Y., Kelmanson, I. V., Bielawski, J. P. & Matz, M. V. Adaptive Evolution of Multicolored Fluorescent Proteins in Reef-Building Corals. *J Mol Evol* **62**, 332–339 (2006).

64. Alieva, N. O. *et al.* Diversity and Evolution of Coral Fluorescent Proteins. *PLoS ONE* **3**, e2680 (2008).

65. Palm, G. J. *et al.* The structural basis for spectral variations in green fluorescent protein. *Nat Struct Mol Biol* **4**, 361–365 (1997).

66. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

67. Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).

68. Ferrada, E. & Wagner, A. Evolutionary Innovations and the Organization of Protein Functions in Genotype Space. *PLoS ONE* **5**, e14172 (2010).

69. Harms, M. J. & Thornton, J. W. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **512**, 203–207 (2014).

70. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Science* **25**, 1204–1218 (2016).

71. Schaper, S., Johnston, I. G. & Louis, A. A. Epistasis can lead to fragmented neutral spaces and contingency in evolution. *Proc. R. Soc. B.* **279**, 1777–1783 (2012).

72. Catalán, P., Arias, C. F., Cuesta, J. A. & Manrubia, S. Adaptive multiscapes: an up-to-date metaphor to visualize molecular adaptation. *Biol Direct* **12**, 7, s13062-017-0178–1 (2017).

73. Gonzalez Somermeyer, L. *et al.* Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* **11**, e75842 (2022).

74. Sikosek, T. & Chan, H. S. Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface.* **11**, 20140419 (2014).

75. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5869–5874 (2006).

76. Wachter, R. Photoconvertible Fluorescent Proteins and the Role of Dynamics in Protein Evolution. *IJMS* **18**, 1792 (2017).

77. Petrović, D., Risso, V. A., Kamerlin, S. C. L. & Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface.* **15**, 20180330 (2018).

78. Zou, T., Risso, V. A., Gavira, J. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Molecular Biology and Evolution* **32**, 132–143 (2015).

79. Campbell, E. C. *et al.* Laboratory evolution of protein conformational dynamics. *Current Opinion in Structural Biology* **50**, 49–57 (2018).

80. Gardner, J. M., Biler, M., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Manipulating Conformational Dynamics To Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catal.* **10**, 4863–4870 (2020).

81. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **324**, 203–207 (2009).

82. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat Chem Biol* **12**, 944–950 (2016).

83. Gurskaya, N. G., Savitsky, A. P., Yanushevich, Y. G., Lukyanov, S. A. & Lukyanov, K. A. Color transitions in coral's fluorescent proteins by site-directed mutagenesis. *BMC Biochemistry* (2001).

84. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79–88 (2006).

85. Heim, R. & Tsien, R. Y. Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Current Biology* **6**, 178–182 (1996).

86. Wang, L., Jackson, W. C., Steinbach, P. A. & Tsien, R. Y. Evolution of new nonantibody proteins via iterative somatic hypermutation. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16745–16749 (2004).

87. Heim, R., Prasher, D. C. & Tsien, R. Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 12501–12504 (1994).

88. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat Biotechnol* **22**, 1567–1572 (2004).

89. Nguyen, A. W. & Daugherty, P. S. Evolutionary optimization of fluorescent proteins for intracellular FRET. *Nat Biotechnol* **23**, 355–360 (2005).

90. Nienhaus, K. & Nienhaus, G. U. Chromophore photophysics and dynamics in fluorescent proteins of the GFP family. *J. Phys.: Condens. Matter* **28**, 443001 (2016).

91. Kim, H. *et al.* Acid–Base Catalysis and Crystal Structures of a Least Evolved Ancestral GFP-like Protein Undergoing Green-to-Red Photoconversion. *Biochemistry* **52**, 8048–8059 (2013).

92. Mizuno, H. *et al.* Photo-Induced Peptide Cleavage in the Green-to-Red Conversion of a Fluorescent Protein. *Molecular Cell* **12**, 1051–1058 (2003).

93. Krueger, T. D. *et al.* Dual Illumination Enhances Transformation of an Engineered Green-to-Red Photoconvertible Fluorescent Protein. *Angew. Chem.* **132**, 1661–1669 (2020).

94. Shcherbakova, D. M. *et al.* Bright monomeric near-infrared fluorescent proteins as tags and biosensors for multiscale imaging. *Nat Commun* **7**, 12405 (2016).

95. Krueger, T. D. *et al.* To twist or not to twist: From chromophore structure to dynamics inside engineered photoconvertible and photoswitchable fluorescent proteins. *Protein Science* **32**, (2023).

96. Maillard, J. *et al.* Universal quenching of common fluorescent probes by water and alcohols. *Chem. Sci.* **12**, 1352–1362 (2021).

97. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput Biol* **4**, e1000002 (2008).

98. Kurahashi, R., Sano, S. & Takano, K. Protein Evolution is Potentially Governed by Protein Stability: Directed Evolution of an Esterase from the Hyperthermophilic Archaeon Sulfolobus tokodaii. *J Mol Evol* **86**, 283–292 (2018).

99. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596–604 (2009).

100. Arai, R. Design of helical linkers for fusion proteins and protein-based nanostructures. in *Methods in Enzymology* vol. 647 209–230 (Elsevier, 2021).

101. Castillo-Hair, S. M. *et al.* FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. *ACS Synth. Biol.* **5**, 774–780 (2016).

102. Pace, C. N. [14]Determination and analysis of urea and guanidine hydrochloride denaturation curves. in *Methods in Enzymology* vol. 131 266–280 (Elsevier, 1986).

103. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry* **48**, 545–600 (1997).

104. Weber, G. Energetics of ligand binding to proteins. in *Advances in protein chemistry* vol. 29 1–83 (Elsevier, 1975).

105. Bahar, I., Lezon, T. R., Yang, L.-W. & Eyal, E. Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics* **39**, 23–42 (2010).

106. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–339 (2014).

107. Wei, G., Xi, W., Nussinov, R. & Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **116**, 6516–6551 (2016).

108. Corbella, M., Pinto, G. P. & Kamerlin, S. C. L. Loop dynamics and the evolution of enzyme activity. *Nat Rev Chem* 1–12 (2023) doi:10.1038/s41570-023-00495-w.

109. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **324**, 203–207 (2009).

110. Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences* **106**, 21149–21154 (2009).

111. He, Y., Chen, Y., Alexander, P. A., Bryan, P. N. & Orban, J. Mutational Tipping Points for Switching Protein Folds and Functions. *Structure* **20**, 283–291 (2012).

112. Sikosek, T., Krobath, H. & Chan, H. S. Theoretical Insights into the Biophysics of Protein Bi-stability and Evolutionary Switches. *PLOS Computational Biology* **12**, e1004960 (2016).

113. Damry, A. M. & Jackson, C. J. The evolution and engineering of enzyme activity through tuning conformational landscapes. *Protein Engineering, Design and Selection* **34**, gzab009 (2021).

114. Dishman, A. F. *et al.* Evolution of fold switching in a metamorphic protein. *Science* **371**, 86–90 (2021).

115. Gardner, J. M., Biler, M., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Manipulating Conformational Dynamics To Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catal.* **10**, 4863–4870 (2020).

116. Kaczmarski, J. A. *et al.* Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme. *Nat Commun* **11**, 5945 (2020).

117. Nguyen, V. *et al.* Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **355**, 289–294 (2017).

118. Nixon, C. F. *et al.* Exploring the Evolutionary History of Kinetic Stability in the α-Lytic Protease Family. *Biochemistry* **60**, 170–181 (2021).

119. Okafor, C. D. *et al.* Structural and Dynamics Comparison of Thermostability in Ancient, Modern, and Consensus Elongation Factor Tus. *Structure* **26**, 118-129.e3 (2018).

120. Whitney, D. S., Volkman, B. F. & Prehoda, K. E. Evolution of a Protein Interaction Domain Family by Tuning Conformational Flexibility. *J. Am. Chem. Soc.* **138**, 15150–15156 (2016).

121. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology* **20**, 360–366 (2010).

122. Hochberg, G. K. A. & Thornton, J. W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys* **46**, 247–269 (2017).

123. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072–1080 (2012).

124. Rivoire, O., Reynolds, K. A. & Ranganathan, R. Evolution-Based Functional Decomposition of Proteins. *PLOS Computational Biology* **12**, e1004817 (2016).

125. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods* **11**, 801–807 (2014).

126. Otten, R. *et al.* How directed evolution reshapes energy landscapes in enzymes to boost catalysis. *Science* **370**, 1442–1446 (2020).

127. Arenas, M. Methodologies for Microbial Ancestral Sequence Reconstruction. in *Environmental Microbial Evolution: Methods and Protocols* (ed. Luo, H.) 283–303 (Springer US, 2022). doi:10.1007/978-1-0716-2691-7_14.

128. Liberles, D. A. *Ancestral Sequence Reconstruction*. (OUP Oxford, 2007).

129. Merkl, R. & Sterner, R. Reconstruction of ancestral enzymes. *Perspectives in Science* **9**, 17–23 (2016).

130. Selberg, A. G. A., Gaucher, E. A. & Liberles, D. A. Ancestral Sequence Reconstruction: From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond. *J Mol Evol* **89**, 157–164 (2021).

131. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr Opin Struct Biol* **69**, 131–141 (2021).

132. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

133. Musil, M. *et al.* FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction. *Briefings in Bioinformatics* **22**, bbaa337 (2021).

134. Orlandi, K. N., Phillips, S. R., Sailer, Z. R., Harman, J. L. & Harms, M. J. Topiary: Pruning the manual labor from ancestral sequence reconstruction. *Protein Science* **32**, e4551 (2023).

135. Linus Pauling & Emile Zuckerlandl. Chemical paleogenetics. *Acta Chemica Scandinavica* **17**, S9–S16 (1963).

136. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).

137. Yang, Z., Kumar, S. & Nei, M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650 (1995).

138. Bailleul, G. *et al.* Evolution of enzyme functionality in the flavin-containing monooxygenases. *Nat Commun* **14**, 1042 (2023).

139. Busch, F. *et al.* Ancestral Tryptophan Synthase Reveals Functional Sophistication of Primordial Enzyme Complexes. *Cell Chemical Biology* **23**, 709–715 (2016).

140. Chiang, C.-H. *et al.* Deciphering the evolution of flavin-dependent monooxygenase stereoselectivity using ancestral sequence reconstruction. *Proceedings of the National Academy of Sciences* **120**, e2218248120 (2023).

141. Clifton, B. E. *et al.* Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat Chem Biol* **14**, 542–547 (2018).

142. Furukawa, R., Toma, W., Yamazaki, K. & Akanuma, S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. *Sci Rep* **10**, 15493 (2020).

143. Garcia, A. K., McShea, H., Kolaczkowski, B. & Kaçar, B. Reconstructing the evolutionary history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization. *Geobiology* **18**, 394–411 (2020).

144. Rauwerdink, A. *et al.* Evolution of a Catalytic Mechanism. *Molecular Biology and Evolution* **33**, 971–979 (2016).

145. Gaucher, E. A., Thomson, J. M., Burgan, M. F. & Benner, S. A. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**, 285–288 (2003).

146. Hart, K. M. *et al.* Thermodynamic System Drift in Protein Evolution. *PLOS Biology* **12**, e1001994 (2014).

147. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345**, 86–89 (1990).

148. Lim, S. A., Bolin, E. R. & Marqusee, S. Tracing a protein's folding pathway over evolutionary time using ancestral sequence reconstruction and hydrogen exchange. *eLife* **7**, e38369 (2018).

149. Smock, R. G., Yadid, I., Dym, O., Clarke, J. & Tawfik, D. S. De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell* **164**, 476–486 (2016).

150. Sang, D. *et al.* Ancestral reconstruction reveals mechanisms of ERK regulatory evolution. *eLife* **8**, e38805 (2019).

151. East, N. J., Clifton, B. E., Jackson, C. J. & Kaczmarski, J. A. The role of oligomerization in the optimization of cyclohexadienyl dehydratase conformational dynamics and catalytic activity. *Protein Science* **31**, e4510 (2022).

152. Holinski, A., Heyn, K., Merkl, R. & Sterner, R. Combining ancestral sequence reconstruction with protein design to identify an interface hotspot in a key metabolic enzyme complex. *Proteins: Structure, Function, and Bioinformatics* **85**, 312–321 (2017).

153. Baldwin, M. W. *et al.* Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. *Science* **345**, 929–933 (2014).

154. Howard, C. J. *et al.* Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. *eLife* **3**, e04126 (2014).

155. McKeown, A. N. *et al.* Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58–68 (2014).

156. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Current Opinion in Structural Biology* **47**, 113–122 (2017).

157. Field, S. F. & Matz, M. V. Retracing Evolution of Red Fluorescence in GFP-Like Proteins from Faviina Corals. *Molecular Biology and Evolution* **27**, 225–233 (2010).

158. Kim, H. *et al.* A Hinge Migration Mechanism Unlocks the Evolution of Green-to-Red Photoconversion in GFP-like Proteins. *Structure* **23**, 34–43 (2015).

159. Castiglione, G. M. & Chang, B. S. Functional trade-offs and environmental variation shaped ancient trajectories in the evolution of dim-light vision. *eLife* **7**, e35957 (2018).

160. Van Nynatten, A., Castiglione, G. M., de A. Gutierrez, E., Lovejoy, N. R. & Chang, B. S. W. Recreated Ancestral Opsin Associated with Marine to Freshwater Croaker Invasion Reveals Kinetic and Spectral Adaptation. *Molecular Biology and Evolution* **38**, 2076–2087 (2021).

161. Yokoyama, S. & Radlwimmer, F. B. The molecular genetics and evolution of red and green color vision in vertebrates. *Genetics* **158**, 1697–1710 (2001).

162. Yokoyama, S., Tada, T., Zhang, H. & Britt, L. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences* **105**, 13480–13485 (2008).

163. Chi, P. B. & Liberles, D. A. Selection on protein structure, interaction, and sequence. *Protein Science* **25**, 1168–1178 (2016).

164. Cortez, L. M. *et al.* Probing the origin of prion protein misfolding via reconstruction of ancestral proteins. *Protein Science* **31**, e4477 (2022).

165. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* **5**, 366–375 (2004).

166. Le, S. Q. & Gascuel, O. An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**, 1307–1320 (2008).

167. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* **11**, 367–372 (1996).

168. Le, S. Q. & Gascuel, O. Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial. *Systematic Biology* **59**, 277–287 (2010).

169. Moshe, A. & Pupko, T. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. *Bioinformatics* **35**, 2562–2568 (2019).

170. Groussin, M. *et al.* Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees. *Molecular Biology and Evolution* **32**, 13–22 (2015).

171. Pagel, M., Meade, A. & Barker, D. Bayesian Estimation of Ancestral Character States on Phylogenies. *Systematic Biology* **53**, 673–684 (2004).

172. Straub, K. & Merkl, R. Ancestral Sequence Reconstruction as a Tool for the Elucidation of a Stepwise Evolutionary Adaptation. in *Computational Methods in Protein Evolution* (ed. Sikosek, T.) 171–182 (Springer, 2019). doi:10.1007/978-1-4939-8736-8_9.

173. Mallik, S., Tawfik, D. S. & Levy, E. D. How gene duplication diversifies the landscape of protein oligomeric state and function. *Current Opinion in Genetics & Development* **76**, 101966 (2022).

174. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012).

175. Pillai, A. S. *et al.* Origin of complexity in haemoglobin evolution. *Nature* **581**, 480–485 (2020).

176. Schulz, L. *et al.* Evolution of increased complexity and specificity at the dawn of form I Rubiscos. *Science* **378**, 155–160 (2022).

177. Lee, J. & Blaber, M. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proceedings of the National Academy of Sciences* **108**, 126–130 (2011).

178. Yang, G. *et al.* Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat Chem Biol* **15**, 1120–1128 (2019).

179. Kaltenbach, M. *et al.* Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat Chem Biol* **14**, 548–555 (2018).

180. Joho, Y. *et al.* Ancestral Sequence Reconstruction Identifies Structural Changes Underlying the Evolution of Ideonella sakaiensis PETase and Variants with Improved Stability and Activity. *Biochemistry* **62**, 437–450 (2023).

181. Wilson, C. *et al.* Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347**, 882–886 (2015).

182. Hadzipasic, A. *et al.* Ancient origins of allosteric activation in a Ser-Thr kinase. *Science* **367**, 912–917 (2020).

183. Natarajan, C. *et al.* Evolution and molecular basis of a novel allosteric property of crocodilian hemoglobin. *Current Biology* **33**, 98-108.e4 (2023).

184. Park, Y., Patton, J. E. J., Hochberg, G. K. A. & Thornton, J. W. Comment on "Ancient origins of allosteric activation in a Ser-Thr kinase". *Science* **370**, eabc8301 (2020).

185. Schupfner, M., Straub, K., Busch, F., Merkl, R. & Sterner, R. Analysis of allosteric communication in a multienzyme complex by ancestral sequence reconstruction. *Proceedings of the National Academy of Sciences* **117**, 346–354 (2020).

186. Boucher, J. I., Jacobowitz, J. R., Beckett, B. C., Classen, S. & Theobald, D. L. An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *eLife* **3**, e02304 (2014).

187. Harman, J. L. *et al.* Evolution of multifunctionality through a pleiotropic substitution in the innate immune protein S100A9. *eLife* **9**, e54100 (2020).

188. Harman, J. L. *et al.* Evolution avoids a pathological stabilizing interaction in the immune protein S100A9. *Proceedings of the National Academy of Sciences* **119**, e2208029119 (2022).

189. Harms, M. J. & Thornton, J. W. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **512**, 203–207 (2014).

190. Kumar, A. *et al.* Stability-Mediated Epistasis Restricts Accessible Mutational Pathways in the Functional Evolution of Avian Hemoglobin. *Molecular Biology and Evolution* **34**, 1240–1251 (2017).

191. Tufts, D. M. *et al.* Epistasis Constrains Mutational Pathways of Hemoglobin Adaptation in High-Altitude Pikas. *Molecular Biology and Evolution* **32**, 287–298 (2015).

192. Duan, S. *et al.* Epistatic interactions between neuraminidase mutations facilitated the emergence of the oseltamivir-resistant H1N1 influenza viruses. *Nat Commun* **5**, 5029 (2014).

193. Starr, T. N., Flynn, J. M., Mishra, P., Bolon, D. N. A. & Thornton, J. W. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proceedings of the National Academy of Sciences* **115**, 4453–4458 (2018).

194. Bridgham, J. T., Ortlund, E. A. & Thornton, J. W. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515–519 (2009).

195. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631 (2013).

196. Studer, R. A., Christin, P.-A., Williams, M. A. & Orengo, C. A. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences* **111**, 2223–2228 (2014).

197. Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).

198. Horovitz, A. & Fersht, A. R. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *Journal of Molecular Biology* **214**, 613–617 (1990).

199. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A. R. Strength and co-operativity of contributions of surface salt bridges to protein stability. *Journal of Molecular Biology* **216**, 1031–1044 (1990).

200. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development* **23**, 700–707 (2013).

201. Morrison, A. J., Wonderlick, D. R. & Harms, M. J. Ensemble epistasis: thermodynamic origins of nonadditivity between mutations. *Genetics* **219**, iyab105 (2021).

202. Jalal, A. S. B. *et al.* Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. *Cell Reports* **32**, 107928 (2020).

203. Akanuma, S. *et al.* Experimental evidence for the thermophilicity of ancestral life. *Proceedings of the National Academy of Sciences* **110**, 11067–11072 (2013).

204. Butzin, N. C. *et al.* Reconstructed Ancestral Myo-Inositol-3-Phosphate Synthases Indicate That Ancestors of the Thermococcales and Thermotoga Species Were More Thermophilic than Their Descendants. *PLOS ONE* **8**, e84300 (2013).

205. Garcia, A. K., Schopf, J. W., Yokobori, S., Akanuma, S. & Yamagishi, A. Reconstructed ancestral enzymes suggest long-term cooling of Earth's photic zone since the Archean. *Proceedings of the National Academy of Sciences* **114**, 4619–4624 (2017).

206. Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704–707 (2008).

207. Iwabata, H., Watanabe, K., Ohkuri, T., Yokobori, S. & Yamagishi, A. Thermostability of ancestral mutants of Caldococcus noboribetus isocitrate dehydrogenase. *FEMS Microbiology Letters* **243**, 393–398 (2005).

208. Miyazaki, J. *et al.* Ancestral Residues Stabilizing 3-Isopropylmalate Dehydrogenase of an Extreme Thermophile: Experimental Evidence Supporting the Thermophilic Common Ancestor Hypothesis. *The Journal of Biochemistry* **129**, 777–782 (2001).

209. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β-Lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902 (2013).

210. Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M. & Harmer, N. Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun Biol* **2**, 1–12 (2019).

211. Emond, S. *et al.* Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nat Commun* **11**, 3469 (2020).

212. Gomez-Fernandez, B. J., Risso, V. A., Rueda, A., Sanchez-Ruiz, J. M. & Alcalde, M. Ancestral Resurrection and Directed Evolution of Fungal Mesozoic Laccases. *Applied and Environmental Microbiology* **86**, e00778-20 (2020).

213. Hobbs, H. T. *et al.* Saturation mutagenesis of a predicted ancestral Syk-family kinase. *Protein Science* **31**, e4411 (2022).

214. Risso, V. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Biotechnological and protein-engineering implications of ancestral protein resurrection. *Current Opinion in Structural Biology* **51**, 106–115 (2018).

215. Schenkmayerova, A. *et al.* Engineering the protein dynamics of an ancestral luciferase. *Nat Commun* **12**, 3616 (2021).

216. Trudeau, D. L. & Tawfik, D. S. Protein engineers turned evolutionists—the quest for the optimal starting point. *Current Opinion in Biotechnology* **60**, 46–52 (2019).

217. Scotese, C. R., Song, H., Mills, B. J. W. & van der Meer, D. G. Phanerozoic paleotemperatures: The earth's changing climate during the last 540 million years. *Earth-Science Reviews* **215**, 103503 (2021).

218. Woese, C. R. Bacterial evolution. *Microbiological Reviews* **51**, 221–271 (1987).

219. Garcia, A. K. & Kaçar, B. How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radical Biology and Medicine* **140**, 260–269 (2019).

220. Wheeler, L. C., Lim, S. A., Marqusee, S. & Harms, M. J. The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology* **38**, 37–43 (2016).

221. Sternke, M., Tripp, K. W. & Barrick, D. Chapter Seven - The use of consensus sequence information to engineer stability and activity in proteins. in *Methods in Enzymology* (ed. Tawfik, D. S.) vol. 643 149–179 (Academic Press, 2020).

222. Susko, E. & Roger, A. J. Problems With Estimation of Ancestral Frequencies Under Stationary Models. *Systematic Biology* **62**, 330–338 (2013).

223. Trudeau, D. L., Kaltenbach, M. & Tawfik, D. S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Molecular Biology and Evolution* **33**, 2633–2641 (2016).

224. Williams, P. D., Pollock, D. D., Blackburne, B. P. & Goldstein, R. A. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. *PLOS Computational Biology* **2**, e69 (2006).

225. Pollock, D. D., Thiltgen, G. & Goldstein, R. A. Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences* **109**, E1352–E1359 (2012).

226. Risso, V. A. *et al.* Mutational Studies on Resurrected Ancestral Proteins Reveal Conservation of Site-Specific Amino Acid Preferences throughout Evolutionary History. *Molecular Biology and Evolution* **32**, 440–455 (2015).

227. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins: Structure, Function, and Bioinformatics* **46**, 105–109 (2002).

228. Hilton, S. K. & Bloom, J. D. Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence. *Virus Evolution* **4**, vey033 (2018).

229. Ho, L. S. T. & Susko, E. Ancestral state reconstruction with large numbers of sequences and edge-length estimation. *J. Math. Biol.* **84**, 21 (2022).

230. Arenas, M. & Bastolla, U. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. *Methods in Ecology and Evolution* **11**, 248–257 (2020).

231. Wheeler, L. C., Anderson, J. A., Morrison, A. J., Wong, C. E. & Harms, M. J. Conservation of Specificity in Two Low-Specificity Proteins. *Biochemistry* **57**, 684–695 (2018).

232. Del Amparo, R. & Arenas, M. Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Molecular Biology and Evolution* **39**, msac144 (2022).

233. Holland, B. R., Ketelaar-Jones, S., O'Mara, A. R., Woodhams, M. D. & Jordan, G. J. Accuracy of ancestral state reconstruction for non-neutral traits. *Sci Rep* **10**, 7644 (2020).